Summary

Introduction

Current approaches to object recognition make essential use of machine learning methods. To improve their performance we need larger datasets ,but it has only recently become possible to collect labeled datasets with millions of images . The new larger datasets include LabelMe and lmageNet . AlexNet was a game changer in deep learning especially in image classification. the immense complexity of the object recognition task means that this problem cannot be specified even by a dataset as large as lmageNet, so our model should also have lots of prior knowledge to compensate for all the data we don't have.
Convolutional neural networks constitute one such class of models. they trained one of the largest convolutional neural networks to date on the subsets of lmageNet used in the ILSVRC-2010 and ILSVRC-2012 competitions and achieved by far the best results ever reported on these datasets.

Dataset

lmageNet is a dataset of over 15 million labeled high-resolution images belonging to roughly 22,000
categories. lmageNet consists of variable-resolution images, while system requires a constant input dimensions, so images were down-sampled to a fixed resolution of 256 × 256 .Given a rectangular image, they are first rescaled such that the shorter side was of length 256, and then cropped out the central 256 × 256 patch from the resulting image. images was not pre-processed in any other way, except for subtracting the mean activity over the training set from each pixel

Architectural Details

AlexNet contains eight learned layers five convolutional and three fully-connected layers

Use of ReLU non linearity :

In terms of training time with gradient descent, saturating nonlinearities like f(x) = tanh(x) are much slower than the non-saturating nonlinearity f(x) = max(0,x). Deep convolutional neural networks with ReLUs train several times faster than

their equivalents with tanh units. as faster learning has a great influence on the performance of large models trained on large datasets they have used ReLU in their network

Training on Multiple GPUs :

A single GTX 580 GPU has only 3GB of memory, which limits the maximum size of the networks
that can be trained on it. Therefore net was spread across two GPUs. The parallelization scheme that they have employed essentially puts half of the kernels (or neurons) on each GPU, with one
additional trick: the GPUs communicate only in certain layers. This means that, for example, the
kernels of layer 3 take input from all kernel maps in layer 2. However, kernels in layer 4 take input
only from those kernel maps in layer 3 which reside on the same GPU.

Local Response Normalization :

ReLUs have the desirable property that they do not require input normalization to prevent them
from saturating . However they still found that the following local normalization scheme aids
generalization. Response normalization reduces the top-1 and top-5 error rates by 1.4% and 1.2% respectively. We also verified the effectiveness of this scheme on the CIFAR-10 dataset: a four-layer CNN achieved a 13% test error rate without normalization and 11% with normalization.

Pooling :

Overlapping pooling has been used in network not traditional pooling . This scheme reduces the top-1 and top-5 error rates by 0.4% and 0.3%, respectively, as compared with the non-overlapping scheme . they observed that during training that models with overlapping pooling they find it slightly more difficult to overfit

Overall Architecture contains eight layers with weights the first five are convolutional and the remaining three are fully connected. The output of the last fully-connected layer is fed to a 1000-way softmax which produces a distribution over the 1000 class labels. The kernels of the second, fourth, and fifth

convolutional layers are connected only to those kernel maps in the previous layer which reside on the same GPU. The kernels of the third convolutional layer are connected to all kernel maps in the second layer. The neurons in the fully connected layers are connected to all neurons in the previous layer. Response-normalization layers follow the first and second convolutional layers. Max-pooling layers, of the kind described in follow both response-normalization layers as well as the fifth convolutional layer. The ReLU non-linearity is applied to the output of every convolutional and fully-connected layer. The third, fourth, and fifth convolutional layers are connected to one another without any intervening pooling or normalization layers. The fully-connected layers have 4096 neurons each.

Overfitting

This neural network architecture has 60 million parameters. size of network made overfitting a significant problem.The easiest and most common method to reduce overfitting on image data is to do Data Augmentation , data augmentation schemes which are used here are in effect computationally free .The first form of data augmentation consists of generating image translations and horizontal reflections. this is done by extracting random 224 × 224 patches (and their horizontal reflections) from the 256 × 256 images and training network on these extracted patches . This increases the size of training set by a factor of 2048. At test time, the network makes a prediction by extracting five 224 224 patches (the four corner patches and the center patch) as well as their horizontal reflections (hence ten patches in all), and averaging the predictions made by the network’s softmax layer on the ten patches. The second form of data augmentation consists of altering the intensities of the RGB channels in training images. Specifically, we perform PCA on the set of RGB pixel values throughout the ImageNet training set. This scheme approximately captures an important property of natural images, namely, that object identity is invariant to changes in the intensity and color of the illumination. This scheme reduces the top-1 error rate by over 1%.

Dropout

This technique reduces complex co-adaptations of neurons, since a neuron cannot rely on the presence of particular other neurons It is, therefore, forced to learn more robust features that are useful in conjunction with many different random subsets of the other neurons. At test time, we use all the neurons but

multiply their outputs by 0.5 . they used dropout in the first two fully-connected layers. Without dropout, our network exhibits substantial overfitting. Dropout roughly doubles the number of iterations required to converge

Learning

they trained models using stochastic gradient descent with a batch size of 128 examples, momentum of 0.9, and weight decay of 0.0005. they found that this small amount of weight decay was important for the model to learn. weights in each layer were initialized with zero-mean Gaussian distribution with standard deviation 0.01. neuron biases in the second, fourth, and fifth layers, as well as in the fully-connected layers with the constant 1. This initialization accelerates the early stages of learning by providing the ReLUs with positive inputs .equal learning rate was used for all layers and it was reduced by a factor of 10 when it stopped improving validation error rate . trained the network for roughly 90 cycles through the training set of 1.2 million images, which took five to six days on two NVIDIA GTX 580 3GB GPUs.

Results and Impact

results has shown that a large deep convolutional neural network is capable of achieving record
breaking results on a highly challenging dataset using purely supervised learning. it was observed that network’s performance degrades if a single convolutional layer is removed . so depth is very important.