ImageNet Classification with Deep Convolutional Neural Networks

Abstract

model
60 million parameters
650,000 neurons
five convolutional layers, some of which are followed by max-pooling layers,
and three fully-connected layers with a final 1000-way softmax.

faster training
non-saturating neurons
a very efficient GPU implementation of the convolution operation

reduce overfitting in fully-connected layers
dropout

Discussion

depth really is important

Introduction

기존 개선법
collect larger datasets,
learn more powerful models
use better techniques for preventing overfitting.

Until recently, datasets of labeled images were relatively small

Simple recognition tasks에 대해서는 좋다
objects in realistic settings은 much larger training sets이 필요

standard feedforward neural networks with similarly-sized layers,보다 CNN은 much fewer connections and parameters and so they are easier to train 대신 이론적 결과는 좀 안 좋을 수 있다

they have still been prohibitively expensive to apply in large scale to high-resolution images
current GPUs are powerful enough to facilitate the training of interestingly-large CNNs
recent datasets

new and unusual features for boosting -> section 3
overfitting prevention -> section 4

Dataset

ImageNet
down-sampled the images to a fixed resolution of 256 × 256
We did not pre-process the images
그러나 subtracting the mean activity over the training set from each pixel는 함

Architecture

5 convs and 3 fc

특이한 점들
1 ReLU Nonlinearity

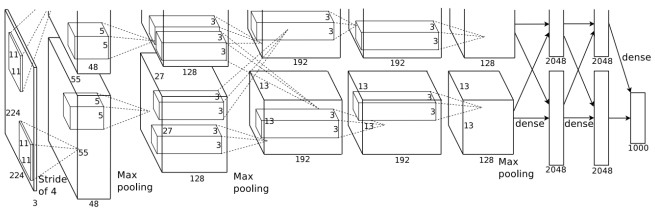
ReLU가 tanh보다 더 빠르더라

2 Multiple GPUs
GPU가 각자 메모리에 바로 접근이 가능해서 병렬화가 좋더라
그래서 GPU 2개에 나눠 학습시켰다

3 Local Response Normalization
ReLU는 saturating을 막기 위해 input normalization을 할 필요는 없다
그치만 local normalization scheme을 하면 좋다
hyperparameter가 있다
ReLU 이후에 다 적용했다

4 Overlapping Pooling
overfit이 방지되더라

Overall Architecture
5 convs and 3 fc
1000 class



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 fullyconnected 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 Section 3.4, 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 first convolutional layer filters the 224×224×3 input image with 96 kernels of size 11×11×3 with a stride of 4 pixels
second 256 kernels of size 5 × 5 × 48.

The third convolutional layer has 384 kernels of size 3 × 3 × 256 connected to the (normalized, pooled) outputs of the second convolutional layer.
The fourth convolutional layer has 384 kernels of size 3 × 3 × 192
the fifth convolutional layer has 256 kernels of size 3 × 3 × 192.
The fully-connected layers have 4096 neurons each.

Reducing Overfitting

60 million parameters
overfit은 안나오기 힘들다

1 Data Augmentation
Enlarge하라
디스크에 저장될 필요 없도록 적은 연산으로 가능
CPU가 만들도록 하였다 - 계산적으로 무료

1 translations & horizontal reflections
원본 & 대칭
256 x 256에서 224 x 224 패치 랜덤 뜯어서 학습

test에서는 가운데와 4 코너 + 대칭으로 10개 뽑아서 prediction average

2 RGB pixel에 값 더하기

2 Dropout
probability 0.5
first two fully-connected layers
roughly doubles the number of iterations required to converge

test에서는 0.5를 곱한다

Details of Learning

Stochastic Gradient Descent
Batch Size 128
Momentum 0.9
Weight Decay 0.0005 - 꽤나 중요

초기 weight initialization - 정규분포 (표준편차 0.01)

initialized the neuron biases in the second, fourth, and fifth convolutional layers, as well as in the fully-connected hidden layers, with the constant 1
accelerates the early stages of learning by providing the ReLUs with positive inputs.

We initialized the neuron biases in the remaining layers with the constant 0.

equal learning rate
heuristic
0.01 and divide 10 three times

roughly 90 cycles through the training set of 1.2 million images
five to six days on two NVIDIA GTX 580 3GB GPUs

Results

ILSVRC-2010

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

ILSVRC-2012

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%**

1 Qualitative Evaluations
GPU간의 Specialization이 일어나더라 (Color-Specific / Color-Agnostic)
Off-Center Objecte도 잘 잡더라
대부분의 top-5 라벨이 옳더라

Euclidean Distance
auto-encoder 이용