ImageNet Classificdation with Deep CNN

Author : Jaeseok Huh @ Hanyang University
Date : Jan 29, 2018

Comment

- This paper is focusing mainly on overfitting problem which was problametic at that time
- As the authors predicted, rapid improvement on a GPU has enhanced the research.

Abstract

- LSVRC-2010, 1.2M img / ILSVRC-2012 top-5 15.3% (2nd 26.2%)
- top-1 37.5% / top-5 17.0%
- 60M param, 65K neuron, 5 NN
- non-saturating neuron, dropout

1. Introduction

- CNN have much fewer connections and parameters but theoretically it is likely to be only slightly worse
- Still suffering from overfitting, several tech. used to avoid it – *we don't any longer recently*
- 5 CNN, 3 FC – the depth of CNN is really important
- Trained 5-6 days on two GTX 580 3GB *(equivalent to 2-3 days on GTX 1060 6GB)*

2. Dataset

- centered patch of 256 x 256 from ImageNet
- mean subtraction

3. Architectdure

3.1 ReLU Nonlinearity

- non-saturating nonlinearity $f(x) = max(0, x)$ reaches 25% error as 6 times fast as *tanh*

- ***recently it is widely used to avoid vanish gradient problem***

3.2 Training on Multiple GPUs

two GPUs which communicate only in certain layers

3.3 Local Response Normalization

After applying ReLU, $a^i_{x,y}$ – activity of neuron computed by applying kenrel i at pos (x, y) $b^i_{x,y}$ – response-normalized activity,

$$b^i_{x,y} = a^i_{x,y} / \left(k + \alpha \sum_{j \in kernel} (a^i_{x,y})^2 \right)^{\beta}$$

empirically $k = 2, n = 5, \alpha = 10^{-4}, \beta = 0.75$
contributed to reduce top-5 error by 1.2%

3.4 Overlapping Pooling

stride $s <$ width z
reduces top-5 error by 0.3%

3.5 Overall Architecture

5 CNN - 3FC - Softmax
224x224x3
96 11x11x3 (stride 4)
max pooling
55x55x96 (separated by 2GPUs)
256 5x5x48
max pooling
27x27x256 (separated by 2GPUs)
384 3x3x256
3x3x384 (separated by 2GPUs)
13x13x256
384 3x3x192
13x13x384 (separated by 2GPUs)
256 3x3x192
13x13x256 (separated by 2GPUs)
max pooling

5 CNN - 3FC - Softmax
FC
FC
FC
SoftMax

4. Reducing Overfitting

4.1 Data Augmentation

- Exploited background CPU to produce transformed images from the original image
- For larger dataset, extracted 2048 random 224x224 patches per image, try 5x2(horizontal reflections) random patches from test set and average them after softmax
- PCA on RGB color space

4.2 Dropout

- On two FC layers
- Multiplied by 0.5 at test time – *we typically double it at training time rather than test time for real-time processing*
- Really helpful to avoid overfitting whereas converges slower 2x

5. Details of learning

- SGD with batch size of 128, momentum 0.9 , decay 0.00005
- weights init - $N(0, 0.1^2)$, biases - 0
- l.r - 0.01, divided by 10 whenever val. error remains stationary (3 times performed)
- 5-6 days with two GTX 580

6. Results

6.1 Qualitative Evaluations

GPU 1 appears to be color-agnostic, GPU 2 appears to be largely color-specific. Seems plausible as comparing L2 distances

7. Discussion

- **Deeper** network, better performance
- May expect far better preformance with unsupervised pre-training which was not used to simplify their paper
- Expects further study with infero-temporal gyrus and video sequence beyond merely static images