ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton

Presented by Baotuan Nguyen and Markham Anderson
ECS 289G, 1.18.2018

*Slides adapted from Tugce Tasci and Kyunghee Kim for CS231B at Stanford University
ImageNet

- Over 15M labeled high resolution images
- Roughly 22K categories
- Collected from web and labeled by Amazon Mechanical Turk

http://image-net.org/
ILSVRC

• Annual competition of image classification at large scale
• 1.2M images in 1K categories
• Classification: make 5 guesses about the image label
ILSVRC

ImageNet Classification error throughout years and groups

Architecture

5 Convolutional Layers

3 Fully Connected Layers

1000-way softmax
SuperVision (SV)

Image classification with deep convolutional neural networks

- 7 hidden “weight” layers
- 650K neurons
- 60M parameters
- 630M connections

Why Convolutions for Images?

- Images are very large inputs, (i.e. 200x200 = 40K pixels) not scalable for fully connected input to a traditional NN.
- Convolutions are small filters (i.e. 3x3 = 9 parameters) that we can learn and apply over portions of an image.
Hierarchical representation

Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]
Why Does This Work? Pixel Locality!
Architecture

RELU Nonlinearity

- Standard way to model a neuron
  \[ f(x) = \tanh(x) \quad \text{or} \quad f(x) = \left(1 + e^{-x}\right)^{-1} \]
  Very slow to train

- Non-saturating nonlinearity (RELU)
  \[ f(x) = \max(0, x) \]
  Quick to train
The Vanishing Gradient Problem
RELU to the Rescue
Architecture

RELU Nonlinearity

A 4 layer CNN with ReLUs (solid line) converges six times faster than an equivalent network with tanh neurons (dashed line) on CIFAR-10 dataset.
Architecture

Training on Multiple GPUs

GPU #1

intra-GPU connections

GPU #2

inter-GPU connections
96 Convolutional Kernels

- 11 x 11 x 3 size kernels.
- top 48 kernels on GPU 1: color-agnostic
- bottom 48 kernels on GPU 2: color-specific.
Artificially Enlarging the Dataset

- 60 million parameters
- 1.2 million training images
Data Augmentation: Altering RGB Intensities

“Object identity is invariant to changes in intensity and color of the illumination.”

\[ [I^R_{xy}, I^G_{xy}, I^B_{xy}] + [p_1, p_2, p_3][\alpha_1 \lambda_1, \alpha_2 \lambda_2, \alpha_3 \lambda_3] \]

\( p = \text{eigenvector} \)
\( \lambda = \text{eigenvalue} \)
\( \alpha \sim N(0,0.1) \)

PC: https://photodune.net/item/grand-canal-at-night-venice/4823526
Data Augmentation: Multiple Patch Extraction

Downsample and crop to 256x256

But input layer is 224x224
Data Augmentation: Multiple Patch Extraction
Data Augmentation: Multiple Patch Extraction

- Downsample and crop to 256x256
  - Extract patches of 224x224
  - Horizontal reflection
  - => 32 x 32 x 2 = 2048

60 million parameters vs 108 million images

PC: Presentation by Tugce Tasci, Kyunghee Kim 05/18/2015
Data Augmentation: Other Label-Preserving Transformations

- rotation
- scaling
- translation


- horizontal shearing
- elastic deformations

Data Augmentation: Other Label-Preserving Transformations

Displacement fields with smoothing (bilinear interpolation)

Dropout

PC: Srivastava et al.
## Datasets

<table>
<thead>
<tr>
<th></th>
<th>images</th>
<th>categories</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ImageNet 2009</strong></td>
<td>8.9 million</td>
<td>22000</td>
</tr>
<tr>
<td>†ImageNet 2011</td>
<td>15.0 million</td>
<td>10184</td>
</tr>
<tr>
<td><em>ILSVRC 2010</em></td>
<td>1.2 million</td>
<td>1000</td>
</tr>
<tr>
<td>‡ILSVRC 2012</td>
<td>“”</td>
<td>“”</td>
</tr>
</tbody>
</table>

*primary reporting
†pretraining
‡competition entry
Evaluating Performance: Error

2010

- Winner
- Best Published
- Writer’s w/out 10
- Writer’s mean 10

2012

- Vanilla
- 5 Nets
- 6th Conv, PT
- 6th Conv, PT, 7 Nets
Evaluating Performance
Evaluating Performance: Error
Evaluating Performance: Error

2009
- 10,184 categories
- 8.9 million images

2010, 2012
- 1000 categories
- 1.8 million images
Evaluating Performance: Nearest Neighbours
Autoencoders

Original mushroom

Encode

Compressed Data

Decode

Learned representation

PC: https://medium.com/@curiously/credit-card-fraud-detection-using-autoencoders-in-keras-tensorflow-for-hackers-part-vii-20e0c85301bd
Unsupervised Learning Without Reconstruction

[Hinton & Salakhutdinov, Science 2009]

PC: https://ucdavis.box.com/s/xf5wvae9wxxh159pkqfs0bvr065umh47
Unsupervised Learning Without Reconstruction

Data prediction

Some data

Other data

see also [Vincent et al., 2008]

PC: https://ucdavis.box.com/s/xf5wvae9wxxh159pkgfs0bvr065umh47
Unsupervised Learning Without Reconstruction

PC: https://ucdavis.box.com/s/xf5wvae9wxxh159pkqfs0bvr065umh47
Strengths

- First attempt at deep learning for ILSVRC.
- Paved the path for state of the art in computer vision.
Weaknesses

- Several parts were not well explained, making it difficult for first timers.
- Ex. Local Response Normalization

\[ b^i_{x,y} = a^i_{x,y} / \left( k + \alpha \sum_{j=\max(0,i-n/2)}^{\min(N-1,i+n/2)} (a^j_{x,y})^2 \right)^\beta \]
Discussion

• Depth is really important.
  removing a single convolutional layer degrades the performance.

  K. Simonyan, A. Zisserman.  

  ➔ 16-layer model, 19-layer model. 7.3% top-5 test error on ILSVRC-2012
Going deep and deeper...
Questions?