Convolutional Neural Networks
An Overview

Guilherme Folego
2016-10-27
Objectives

What is a Convolutional Neural Network?

What is it good for?

Why now?
Neural Network

input layer

hidden layer 1  hidden layer 2

output layer
Convolutional Neural Network
Convolutional Neural Network
Convolutional Neural Network
Convolutional Neural Network
**LeNet**


Google Scholar: Cited by 1846
LeNet

Highlights

- “In summary, the network has 1,256 units, 64,660 connections, and 9,760 independent parameters.”
- “... our training times were ‘only’ 3 days”
- “We used an off-the-shelf board that contains 256 kbytes of local memory and ... 25 MFLOPS”
- “This work points out the necessity of having flexible ‘network design’ software tools that ease the design of complex, specialized network architectures”
LeNet

10 output units

layer H3
30 hidden units

layer H2
12 x 16 = 192 hidden units

layer H1
12 x 64 = 768 hidden units

256 input units

fully connected
~ 300 links

fully connected
~ 6000 links

~ 40,000 links from 12 kernels
5 x 5 x 8

~ 20,000 links from 12 kernels
5 x 5
LeNet-5


Google Scholar: Cited by 5964
LeNet-5

Highlights

- Deployed commercially, reading “several million checks per day” (about 15% of all checks in the USA at the time)
- Introduced LeNet-5, arguably the most used CNN for teaching the subject or demonstrating a framework
- “Database: the Modified NIST set” (now known as MNIST, with about 60,000 images)
LeNet-5
“AI winter” for neural nets in the 90’s
Around 2006, some papers on CNN started emerging ...

- CIFAR & “The Deep Learning Conspiracy”
- LeCun, Y., Bengio, Y., and Hinton, G. E.


Google Scholar: Cited by 2964
And the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) started in 2010

- “ImageNet is a dataset of over 15 million labeled high-resolution images belonging to roughly 22,000 categories.”
- “ILSVRC uses a subset of ImageNet with roughly 1,000 images in each of 1,000 categories. In all, there are roughly 1.2 million training images, 50,000 validation images, and 150,000 testing images.”

Google Scholar: Cited by 7153
Krizhevsky (SuperVision)

Highlights

● This paper completely changed the scenario
  ○ The first deep convolutional neural network entry in ILSVRC
  ○ Nearly half the error rate of the second-best entry
    ■ 15.3% vs. 26.2%
● Network named SuperVision
● Code released: cuda-convnet
Krizhevsky (SuperVision)

Highlights

- Network’s size is limited by the amount of memory available
- Between five and six days to train on two GTX 580 3GB GPUs [1,581,100 MFLOPS]
- “All of our experiments suggest that our results can be improved simply by waiting for faster GPUs and bigger datasets to become available.”
Krizhevsky (SuperVision)
Krizhevsky (SuperVision)
Krizhevsky (SuperVision)
Krizhevsky (SuperVision)
The Deep Learning “Computer Vision Recipe”

Big Data: ImageNet + Deep Convolutional Neural Network + Backprop on GPU = Learned Weights

Google Scholar: Cited by 943
 Highlights

● Improved on previous results
  ○ Winner of the localization task
  ○ Very competitive results on the detection and classification tasks
● Network named OverFeat
● Code released
● Network weights released!
### Table 4: Number of parameters and connections for different models.

<table>
<thead>
<tr>
<th>model</th>
<th># parameters (in millions)</th>
<th># connections (in millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Krizhevsky</td>
<td>60</td>
<td>-</td>
</tr>
<tr>
<td>fast</td>
<td>145</td>
<td>2810</td>
</tr>
<tr>
<td>accurate</td>
<td>144</td>
<td>5369</td>
</tr>
</tbody>
</table>

*OverFeat*
OverFeat

<table>
<thead>
<tr>
<th>Model</th>
<th>Pre-training</th>
<th>7 big models</th>
<th>7 fast models</th>
<th>VGG</th>
<th>7 models + ImageNet11</th>
<th>5 models</th>
<th>Top 5 error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clarifai</td>
<td>11.2%</td>
<td>11.7%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OverFeat</td>
<td></td>
<td>13.6%</td>
<td></td>
<td>14.2%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VGG</td>
<td></td>
<td></td>
<td></td>
<td>15.2%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SuperVision</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>15.3%</td>
<td>16.4%</td>
<td></td>
</tr>
</tbody>
</table>

Top 5 error rate

- ILSVRC12
- ILSVRC13
- Post competition
Transfer Learning


Google Scholar: Cited by 728
Transfer Learning

Highlights

- “The results are achieved using a linear SVM classifier (or L2 distance in case of retrieval) applied to a feature representation of size 4096 extracted from a layer in the net.”
Transfer Learning

Highlights

● “The results strongly suggest that features obtained from deep learning with convolutional nets should be the primary candidate in most visual recognition tasks.”
Transfer Learning

Google Scholar: Cited by 32

Micael Cabrera Carvalho’s dissertation
http://www.bibliotecadigital.unicamp.br/document/?code=000956410

Google Scholar: Cited by 2162
VGG

Highlights

● Improved on previous results
  ○ First place in the localization task
  ○ Second place in the classification task
● Network named VGG
● Network architecture is very uniform
● Code based on Caffe framework
● Network weights released!
VGG

Highlights

● “On a system equipped with four NVIDIA Titan Black GPUs [5,120,600 MFLOPS], training a single net took 2–3 weeks depending on the architecture.”
## VGG

<table>
<thead>
<tr>
<th>ConvNet Configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Diagram of VGG" /></td>
</tr>
</tbody>
</table>

### A-LRN

<table>
<thead>
<tr>
<th>11 weight layers</th>
<th>11 weight layers</th>
<th>13 weight layers</th>
<th>16 weight layers</th>
<th>16 weight layers</th>
<th>19 weight layers</th>
</tr>
</thead>
<tbody>
<tr>
<td>conv3-64</td>
<td>conv3-64</td>
<td>conv3-64</td>
<td>conv3-64</td>
<td>conv3-64</td>
<td>conv3-64</td>
</tr>
<tr>
<td>LRN</td>
<td><strong>conv3-64</strong></td>
<td>conv3-64</td>
<td>conv3-64</td>
<td>conv3-64</td>
<td>conv3-64</td>
</tr>
</tbody>
</table>

### Maxpool

<table>
<thead>
<tr>
<th>conv3-128</th>
<th>conv3-128</th>
<th>conv3-128</th>
<th>conv3-128</th>
<th>conv3-128</th>
<th>conv3-128</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>conv3-128</strong></td>
<td><strong>conv3-128</strong></td>
<td>conv3-128</td>
<td>conv3-128</td>
<td>conv3-128</td>
<td>conv3-128</td>
</tr>
</tbody>
</table>

### Maxpool

<table>
<thead>
<tr>
<th>conv3-256</th>
<th>conv3-256</th>
<th>conv3-256</th>
<th>conv3-256</th>
<th>conv3-256</th>
<th>conv3-256</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>conv3-256</strong></td>
<td><strong>conv3-256</strong></td>
<td>conv3-256</td>
<td>conv3-256</td>
<td>conv3-256</td>
<td>conv3-256</td>
</tr>
</tbody>
</table>

### Maxpool

<table>
<thead>
<tr>
<th>conv3-512</th>
<th>conv3-512</th>
<th>conv3-512</th>
<th>conv3-512</th>
<th>conv3-512</th>
<th>conv3-512</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>conv3-512</strong></td>
<td><strong>conv3-512</strong></td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
</tr>
</tbody>
</table>

### Maxpool

<table>
<thead>
<tr>
<th>conv3-512</th>
<th>conv3-512</th>
<th>conv3-512</th>
<th>conv3-512</th>
<th>conv3-512</th>
<th>conv3-512</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>conv3-512</strong></td>
<td><strong>conv3-512</strong></td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
</tr>
</tbody>
</table>

### Maxpool

<table>
<thead>
<tr>
<th>conv3-512</th>
<th>conv3-512</th>
<th>conv3-512</th>
<th>conv3-512</th>
<th>conv3-512</th>
<th>conv3-512</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>conv3-512</strong></td>
<td><strong>conv3-512</strong></td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
</tr>
</tbody>
</table>

### FC

<table>
<thead>
<tr>
<th>4096</th>
<th>4096</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>FC-4096</td>
<td>FC-4096</td>
<td>FC-1000</td>
</tr>
</tbody>
</table>

### Softmax

<table>
<thead>
<tr>
<th>soft-max</th>
</tr>
</thead>
</table>
**Table 2: Number of parameters** (in millions).

<table>
<thead>
<tr>
<th>Network</th>
<th>A, A-LRN</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of parameters</td>
<td>133</td>
<td>133</td>
<td>134</td>
<td>138</td>
<td>144</td>
</tr>
</tbody>
</table>
Artistic Style


Google Scholar: Cited by 91
Artistic Style

Highlights

- Based on VGG network architecture and weights
- “The key finding of this paper is that the representations of content and style in the Convolutional Neural Network are separable. That is, we can manipulate both representations independently to produce new, perceptually meaningful images.”
Artistic Style
Artistic Style
Artistic Style
Driver’s Licence vs. Selfie


Google Scholar: Cited by 1672
GoogLeNet

Highlights

- “... called **GoogLeNet**, a 22 layers deep network, ...”
- “For most of the experiments, the models were designed to keep a computational budget of 1.5 billion multiply-adds at inference time, so that they do not end up to be a purely academic curiosity, but could be put to real world use, even on large datasets, at a reasonable cost.”
- “GoogLeNet networks were trained using the DistBelief distributed machine learning system ...” (lots of CPUs)
GoogLeNet

Highlights
- The first reference is a meme
GoogLeNet
GoogLeNet

Inception
GoogLeNet

Figure 1: Two distinct classes from the 1000 classes of the ILSVRC 2014 classification challenge. Domain knowledge is required to distinguish between these classes.
## GoogLeNet

<table>
<thead>
<tr>
<th>Team</th>
<th>Year</th>
<th>Place</th>
<th>Error (top-5)</th>
<th>Uses external data</th>
</tr>
</thead>
<tbody>
<tr>
<td>SuperVision</td>
<td>2012</td>
<td>1st</td>
<td>16.4%</td>
<td>no</td>
</tr>
<tr>
<td>SuperVision</td>
<td>2012</td>
<td>1st</td>
<td>15.3%</td>
<td>Imagenet 22k</td>
</tr>
<tr>
<td>Clarifai</td>
<td>2013</td>
<td>1st</td>
<td>11.7%</td>
<td>no</td>
</tr>
<tr>
<td>Clarifai</td>
<td>2013</td>
<td>1st</td>
<td>11.2%</td>
<td>Imagenet 22k</td>
</tr>
<tr>
<td>MSRA</td>
<td>2014</td>
<td>3rd</td>
<td>7.35%</td>
<td>no</td>
</tr>
<tr>
<td>VGG</td>
<td>2014</td>
<td>2nd</td>
<td>7.32%</td>
<td>no</td>
</tr>
<tr>
<td>GoogLeNet</td>
<td>2014</td>
<td>1st</td>
<td>6.67%</td>
<td>no</td>
</tr>
</tbody>
</table>
Show and Tell


Google Scholar: Cited by 518
Show and Tell

A group of people shopping at an outdoor market.

There are many vegetables at the fruit stand.
Show and Tell
Show and Tell

A person riding a motorcycle on a dirt road.

Two dogs play in the grass.

A skateboarder does a trick on a ramp.

A dog is jumping to catch a frisbee.

A group of young people playing a game of frisbee.

Two hockey players are fighting over the puck.

A little girl in a pink hat is blowing bubbles.

A refrigerator filled with lots of food and drinks.

A herd of elephants walking across a dry grass field.

A close up of a cat laying on a couch.

A red motorcycle parked on the side of the road.

A yellow school bus parked in a parking lot.

Table: Describes without errors | Describes with minor errors | Somewhat related to the image | Unrelated to the image
ResNet


Google Scholar: Cited by 562
ResNet

Highlights

● First place in ILSVRC 2015 classification, localization, and detection tasks
● “Is learning better networks as easy as stacking more layers?”
● “There exists a solution by construction to the deeper model: the added layers are identity mapping, and the other layers are copied from the learning shallower model.”
  ○ Degradation problem
ResNet

19.6 B FLOPS vs 3.6 B
ResNet

Graphs showing the error (%) against iterations (1e4) for different layer configurations: 18-layer, 34-layer, plain-18, and plain-34. The graphs compare the performance of ResNet and plain models.
**ResNet**

<table>
<thead>
<tr>
<th>method</th>
<th>top-5 err. (test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG [41] (ILSVRC’14)</td>
<td>7.32</td>
</tr>
<tr>
<td>GoogLeNet [44] (ILSVRC’14)</td>
<td>6.66</td>
</tr>
<tr>
<td>VGG [41] (v5)</td>
<td>6.8</td>
</tr>
<tr>
<td>PReLU-net [13]</td>
<td>4.94</td>
</tr>
<tr>
<td>BN-inception [16]</td>
<td>4.82</td>
</tr>
<tr>
<td><strong>ResNet (ILSVRC’15)</strong></td>
<td><strong>3.57</strong></td>
</tr>
</tbody>
</table>

Table 5. Error rates (%) of **ensembles**. The top-5 error is on the test set of ImageNet and reported by the test server.
Object Recognition with Text-to-Speech
Voice Activity Detection

Alzheimer’s Disease

Computer-aided diagnosis for Alzheimer’s disease using 3D convolutional neural networks
Cognitive Computing

Reasoning

Vision

Speech

Dialog

Learning

Signals
Conclusions

Deep Learning is rapidly evolving the machine learning field

Convolutional Neural Networks are key to this advance in the computer vision field

Lots of good data are necessary

Recent technologies are accessible
References

CS231n Convolutional Neural Networks for Visual Recognition
  ●  https://cs231n.github.io/

Deep Learning, Yoshua Bengio, Ian Goodfellow, Aaron Courville, MIT
  ●  http://www.deeplearningbook.org/

A ‘Brief’ History of Neural Nets and Deep Learning