Visualizing and Understanding Convolution Networks

Authors: Mathew Zeiler and Rob Fergus
New York University
Select slides from Hamid Izadinia

Presentation by Jason Driver

10/13/2016
Problem and Contributions

• Greatly improved performance using convolution neural network architecture over the previous state of the art
• No clear understanding why CNN work
• Technique for visualizing individual layers with a Deconvolution Network
• Troubleshoot the output of each layer
• Introduces an image occlusion experiment for spatial understanding
• Optimal architecture
Hierarchical Convolution
Nonlinear operations (max pooling, ReLu)
Convolutional Neural Network
AlexNet
Deconvnet & Convnet
Feature Visualization

corners & edge/color conjunctions
Feature Visualization
Feature Visualization

Layer 4
Object parts
(dog face & bird legs)

Layer 5
Entire object with pose variation
(dogs)
Feature Visualization

Lower Order Features

Higher Order Features
Feature Invariance in Layers 1, 7, 8

Translation

Scale

Rotation

\[ \epsilon_\theta = x_\theta - \tilde{x}_\theta \]

Euclidean distance between feature of transformed and original
Architecture Selection

Layer 1
- Smaller filters (7x7 vs. 11x11) and smaller stride (2 vs. 4)
- 1\textsuperscript{st} Layer: Increased coverage of mid-frequencies (Original is b)

Layer 2
- 2\textsuperscript{nd} Layer: no aliasing, no “dead” features (Original is d)
Occlusion Sensitivity
Correspondence Analysis

- 5th Layer preserves correspondence
- 7th Layer discriminates different classes of dog
## Results

<table>
<thead>
<tr>
<th>Error %</th>
<th>Val Top-1</th>
<th>Val Top-5</th>
<th>Test Top-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Gunji et al., 2012)</td>
<td>-</td>
<td>-</td>
<td>26.2</td>
</tr>
<tr>
<td>(Krizhevsky et al., 2012), 1 convnet</td>
<td>40.7</td>
<td>18.2</td>
<td>--</td>
</tr>
<tr>
<td>(Krizhevsky et al., 2012), 5 convnets</td>
<td>38.1</td>
<td>16.4</td>
<td>16.4</td>
</tr>
<tr>
<td>(Krizhevsky et al., 2012)*, 1 convnets</td>
<td>39.0</td>
<td>16.6</td>
<td>--</td>
</tr>
<tr>
<td>(Krizhevsky et al., 2012)*, 7 convnets</td>
<td>36.7</td>
<td>15.4</td>
<td>15.3</td>
</tr>
<tr>
<td>Our replication of</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Krizhevsky et al., 2012), 1 convnet</td>
<td>40.5</td>
<td>18.1</td>
<td>--</td>
</tr>
<tr>
<td>1 convnet as per Fig. 3</td>
<td>38.4</td>
<td>16.5</td>
<td>--</td>
</tr>
<tr>
<td>5 convnets as per Fig. 3 – (a)</td>
<td>36.7</td>
<td>15.3</td>
<td>15.3</td>
</tr>
<tr>
<td>1 convnet as per Fig. 3 but with layers 3,4,5: 512,1024,512 maps – (b)</td>
<td>37.5</td>
<td>16.0</td>
<td>16.1</td>
</tr>
<tr>
<td>6 convnets, (a) &amp; (b) combined</td>
<td><strong>36.0</strong></td>
<td><strong>14.7</strong></td>
<td><strong>14.8</strong></td>
</tr>
</tbody>
</table>