Understand and Leverage the Internal Representations of Convolutional Neural Networks

Bolei Zhou
MIT
CNN for Image Classification

Large-scale image classification result on ImageNet

Revolution of Depth

ImageNet Classification top-5 error (%)

ILSVRC'16
ILSVRC'15 ResNet
ILSVRC'14 GoogleNet
ILSVRC'14 VGG
ILSVRC'13
ILSVRC'12 AlexNet
ILSVRC'11
ILSVRC'10

2.99
152 layers
28.2
shallow
CNN for General AI

• Alpha Go

• SL policy network is 13 layer-CNN

• Training: 29.4 million positions from 160,000 human professional games.

• CNN beats human professional, can we discover the inside knowledge?

Mastering the game of Go with deep neural networks and tree search
Secret of CNN

Understand and leverage the internal units/representation

final output is a small part of the story
Outline

- Visualizing and annotating the internal units
- Application: weakly supervised localization

Zhou et al. Object Detectors Emerge from Deep Scene CNN. ICLR’15
Zhou et al. Learning Deep Features for Discriminative Localization. CVPR’16
Bau*, and Zhou*, et al. Network Dissection. CVPR’17

Joint work with: David Bau, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba
Object Representations in Computer Vision

dog

vehicle

bird
Object Representations in Computer Vision

dog

vehicle

bird
Object Representations in Computer Vision

**Constellation model**
Weber, Welling & Perona (2000),

**Deformable Part model**

**Bag-of-word model**

**Class-specific graph model**
Kumar, Torr and Zisserman (2005), Felzenszwalb & Huttenlocher (2005)
Object Representations in CNN

Deconvolution


Strong activation image


Back-propagation

Object Representations in CNN

http://people.csail.mit.edu/torralba/research/drawCNN/drawNet.html
A Comparison Study on CNNs

ImageNet CNN for Object Classification

Places CNN for Scene Classification

Same architecture: AlexNet

Zhou et al. Object Detectors Emerge from Deep Scene CNN. ICLR’15
Places: Large-scale Scene Recognition Database

- Places contains 10 million images from ~400 scene categories.
- Data and models are available at [http://places.csail.mit.edu](http://places.csail.mit.edu)

Data-Driven Approach to Visualize CNN

Neuroscientists study brain

200,000 image stimuli of objects and scenes

Zhou et al. Object Detectors Emerge from Deep Scene CNN. ICLR’15
Estimating the Receptive Field of Unit

Estimated receptive fields

Actual size of RF is much smaller than the theoretic size

Zhou et al. Object Detectors Emerge from Deep Scene CNN. ICLR’15
Segmenting Images by Units’ Receptive Fields

Image segmentation using units at different layers:

- conv1
- conv2
- conv4
- conv5

More semantically meaningful

Zhou et al. Object Detectors Emerge from Deep Scene CNN. ICLR’15
Annotating the Semantics of Units

Top ranked segmented images are cropped and sent to Amazon Turk for annotation.

Zhou et al. Object Detectors Emerge from Deep Scene CNN. ICLR’15
Annotating the Semantics of Units

Pool 5, unit 76; Label: ocean; Type: scene; Precision: 93%
Annotating the Semantics of Units

Pool5, unit 13; Label: Lamps; Type: object; Precision: 84%
Annotating the Semantics of Units

Pool5, unit 77; Label: legs; Type: object part; Precision: 96%
Annotating the Semantics of Units

Pool5, unit 112; Label: pool table; Type: object; Precision: 70%
Distribution of Semantic Types at Each Layer
Histogram of Object Detectors in Pool5

ImageNet-CNN

Counts

dog
bird
person
wheel
animal body
flower
ground
head
legs
animal face
animal head
building
car
cat
ceiling
d face
human face
leg
monkey
plant
plants
pot
road
sea
tower
tree
water
window
Histogram of Object Detectors in Pool5

Places-CNN (151/256)
Issue: Manually annotating units is not scalable

Top ranked segmented images are cropped and sent to Amazon Turk for annotation.

Zhou et al. Object Detectors Emerge from Deep Scene CNN. ICLR’15
AlexNet
5 conv layers

11x11 conv, 96, /4, pool/2
5x5 conv, 256, pool/2
3x3 conv, 384
3x3 conv, 384
3x3 conv, 256, pool/2
fc, 4096
fc, 4096
fc, 1000

VGG
16 conv layers

3x3 conv, 64
3x3 conv, 64, pool/2
3x3 conv, 128
3x3 conv, 128, pool/2
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256, pool/2
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512, pool/2
3x3 conv, 512
fc, 4096
fc, 4096
fc, 1000

GoogLeNet
~20 conv layers

ResNet
>100 layers
Solution: Automatic annotation for unit semantics

Corpus of color dataset, texture dataset, shape dataset, object dataset, scene dataset

Bau* and Zhou*, et al. Network Dissection: Quantifying Interpretability of Deep Visual Representations. CVPR’17
Solution: Automatic annotation for unit semantics

Bau* and Zhou*, et al. Network Dissection: Quantifying Interpretability of Deep Visual Representations. CVPR’17
Automatically Annotating Internal Units

Units annotated as concept detectors in the Places-AlexNet
Automatically Annotating Internal Units

Analyzing the effect of training tricks for network interpretability
A zoo of CNN models

<table>
<thead>
<tr>
<th>Training</th>
<th>Network</th>
<th>Dataset or task</th>
</tr>
</thead>
<tbody>
<tr>
<td>N/A</td>
<td>AlexNet</td>
<td>random</td>
</tr>
<tr>
<td>Supervised</td>
<td>AlexNet</td>
<td>ImageNet, Places205, Places365, Hybrid.</td>
</tr>
<tr>
<td></td>
<td>VGG</td>
<td>ImageNet, Places205, Places365, Hybrid.</td>
</tr>
<tr>
<td>Self</td>
<td>AlexNet</td>
<td>context, puzzle, egomotion, tracking, moving, videoorder, audio, crosschannel, colorization, objectcentric.</td>
</tr>
</tbody>
</table>
Supervised CNN on ImageNet/Places

Figure 4: The top ranked tokens identified in the AlexNet, VGG, GoogLeNet, and ResNet on ImageNet and Places365.
Supervised CNN on ImageNet and Places

• Analyzing concept detectors change over layers
Self-supervised CNNs

- Examples of self-supervised training tasks:
  - Context prediction, ICCV’15
  - Solving puzzle, ECCV’16
  - Colorization, ECCV’16 and CVPR’17
  - Predicting video order, ECCV’16
Self-supervised CNNs

- Comparison of supervised CNNs and self-supervised CNNs
Self-supervised CNNs

• Examples of detectors in self-supervised CNNs:
Explanatory factors in deep features
Leveraging the Internal Representations of CNNs

Why CNN makes the prediction?

Prediction from ImageNet-CNN:
Australian terrier:0.75
Why CNN makes the prediction?

Prediction from Places-CNN:
Picnic area: 0.64
Why CNN makes the prediction?

Previous work:
convolutional units as concept detectors at different layers
Simplifying the Network Architecture

Global Average Pooling

conv layers + FC layers + softmax layer
Activation of units as object detectors

Softmax class weights
Class Activation Mapping

- Different classes have different class activation maps
- Top5 predictions: palace, dome, church, altar, monastery
Effect from Removing the FC layers

Classification accuracy drops 2~3%, but with 90% less model parameters

<table>
<thead>
<tr>
<th>Networks</th>
<th>top-1 val. error</th>
<th>top-5 val. error</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGGnet-GAP</td>
<td>33.4</td>
<td>12.2</td>
</tr>
<tr>
<td>GoogLeNet-GAP</td>
<td>35.0</td>
<td>13.2</td>
</tr>
<tr>
<td>AlexNet*-GAP</td>
<td>44.9</td>
<td>20.9</td>
</tr>
<tr>
<td>AlexNet-GAP</td>
<td>51.1</td>
<td>26.3</td>
</tr>
<tr>
<td>GoogLeNet</td>
<td>31.9</td>
<td>11.3</td>
</tr>
<tr>
<td>VGGnet</td>
<td>31.2</td>
<td>11.4</td>
</tr>
<tr>
<td>AlexNet</td>
<td>42.6</td>
<td>19.5</td>
</tr>
<tr>
<td>NIN</td>
<td>41.9</td>
<td>19.6</td>
</tr>
<tr>
<td>GoogLeNet-GMP</td>
<td>35.6</td>
<td>13.9</td>
</tr>
</tbody>
</table>
Weakly-supervised object localization

CNN-GAP is used for object localization, without training with bounding box annotation.

Table 3. Localization error on the ILSVRC test set for various weakly- and fully-supervised methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>supervision</th>
<th>top-5 test error</th>
</tr>
</thead>
<tbody>
<tr>
<td>GoogLeNet-GAP (heuristics)</td>
<td>weakly</td>
<td>37.1</td>
</tr>
<tr>
<td>GoogLeNet-GAP</td>
<td>weakly</td>
<td>42.9</td>
</tr>
<tr>
<td>Backprop [22]</td>
<td>weakly</td>
<td>46.4</td>
</tr>
<tr>
<td>OverFeat [21]</td>
<td>full</td>
<td>29.9</td>
</tr>
<tr>
<td>AlexNet [24]</td>
<td>full</td>
<td>34.2</td>
</tr>
</tbody>
</table>
Localizable Visual Features

Deep CAM feature + linear SVM: localize informative regions

Or any other tasks with any loss functions, like regression, clustering, etc.
Localizable Visual Features

Image captioning using LSTM

Visual question answering

Demo video

https://www.youtube.com/watch?v=fZvOy0VXWA1
http://cnnlocalization.csail.mit.edu
We analyzed the internal representation of CNNs, and leveraged them for weakly-supervised localization.

The papers, the code, and pre-trained models are at

http://places.csail.mit.edu

http://cnnlocalization.csail.mit.edu