Network Dissection:
Quantifying Interpretability of Deep Visual Representations

Bolei Zhou*

MIT

Joint work with David Bau*, Aditya Khosla, Aude Oliva, Antonio Torralba

* Indicates equal contribution
Deep ConvNet for Visual Recognition

2012: AlexNet
- 5 conv. layers
- Error: 15.3%

2014: VGG
- 16 conv. layers
- Error: 8.5%

2015: GoogLeNet
- 22 conv. layers
- Error: 7.8%

2016: ResNet
- >100 conv. layers
- Error: 4.4%
Deep ConvNet for Visual Recognition

2012: AlexNet
5 conv. layers
Error: 15.3%

2014: VGG
16 conv. layers
Error: 8.5%

2015: GoogLeNet
22 conv. layers
Error: 7.8%

2016: ResNet
>100 conv. layers
Error: 4.4%

What have been learned inside?
How to compare the internal representations?
Previous Work on Network Visualization

Deconvolution

Zeiler et al., ECCV 2014.

Back-propagation

Simonyan et al., ICLR 2015

Top activated images

Girshick et al., CVPR 2014

Goal: From Visualization to Interpretation

Top Activated Images

Interpretation: lamp
Score: 0.15

Interpretation: car
Score: 0.02
Approach: Test units for semantic segmentation

Unit 1

Top activated images

Lamp

Intersection over Union (IoU) = 0.12
Network Dissection
Quantifying the interpretability of units through segmentation

Forward pass
Broadly and Densely (Broden) Annotated Dataset

**ADE20K**
- Zhou et al, CVPR’17

**Pascal Context**
- Mottaghi et al, CVPR’14

**Pascal Part**
- Chen et al, CVPR’14

**Open-Surfaces**
- Bell et al, SIGGRAPH’14

**Describable Textures**
- Cimpoi et al, CVPR’14

**Colors**

Total = **63,305** images
- **1,197** visual concepts
AlexNet trained on places

Histogram of object detectors: Detector:81/256, Unique Detector:40 (Units with IoU>0.04)
conv5 unit 79  car (object)  IoU=0.13

conv5 unit 107  road (object)  IoU=0.15

Histogram of object detectors: Detector:81/256, Unique Detector:40 (Units with IoU>0.04)
conv5 unit 144  mountain (object)  IoU=0.13

Histogram of object detectors: Detector:81/256, Unique Detector:40 (Units with IoU>0.04)

conv5 unit 200  mountain (object)  IoU=0.11
Architectures

AlexNet
- 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
- 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
- fc, 4096
- fc, 4096
- fc, 1000

GoogLeNet

ResNet

Supervised Learning

IMAGENET

places

THE SCENE RECOGNITION DATABASE

Self-Supervised Learning

Context prediction, ICCV'15

Solving puzzle, ECCV'16

Colorization, ECCV'16

Audio prediction, ECCV'16
Self-supervised AlexNet

Supervised VGG, GoogLeNet, ResNet

Number of Unique Detectors
Emergence of Interpretable Units during Training

Number of unique detectors

Accuracy on validation set

Training iteration 1

Training iteration 1
Fine-tuning from ImageNet to Places

Unit 8 at conv5 layer

Before fine-tuning
Fine-tuning from ImageNet to Places

Unit 52 at conv5 layer

Before fine-tuning
Fine-tuning from Places to ImageNet

Unit 35 at conv5 layer

Before fine-tuning
Fine-tuning from Places to ImageNet

Unit 103 at conv5 layer

Before fine-tuning
Interpretable Units in DenseNet

- **Horse**
  - Unit 651
  - IoU = 0.151

- **Bus**
  - Unit 740
  - IoU = 0.152

- **Car**
  - Unit 529
  - IoU = 0.150

- **Bed**
  - Unit 953
  - IoU = 0.133

- **Toilet**
  - Unit 1016
  - IoU = 0.125

- **Bakery Shop**
  - Unit 926
  - IoU = 0.167

- **Wheel**
  - Unit 648
  - IoU = 0.165

- **Cobwebbed**
  - Unit 700
  - IoU = 0.327
Conclusion

Welcome to the Poster #11 this afternoon.

Code and more visualizations are at http://netdissect.csail.mit.edu

Living room
Kitchen
Coast
Theater
...

Dissection Report

Histogram of Object Detectors

Histogram of Texture Detectors

unit 79 car, IoU=0.13
unit 107 road, IoU=0.15