Interpreting Deep Visual Representations

Bolei Zhou
MIT

David Bau
Aditya Khosla
Aude Oliva
Antonio Torralba
Background

Convolutional Neural Network (ConvNet)
Many networks

AlexNet

DenseNet

GoogLeNet

VGG Net

ResNet
Why works so well

Upload your image for scene recognition using Places-CNN from MIT.

Predictions:
- **type**: indoor
- **semantic categories**:
  - hotel_room: 0.50, bedroom: 0.47,
When it fails, why is it?

Output: cutting vegetables.
Correct label: gardening

Output: washing dishes.
Correct label: brushing
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?
Work on Network Visualization

Deconvolution
Zeiler et al., ECCV 2014.

Back-propagation
Simonyan et al., ICLR 2015

Feature inversion
Mahendran et al, CVPR 2015

Top activated images
Girshick et al., CVPR 2014
Going From Visualization to Interpretation

Top Activated Images
- **Unit 1**
  - Interpretation: head
  - Score: 0.23

- **Unit 2**
  - Interpretation: lamp
  - Score: 0.15

- **Unit 3**
  - Interpretation: car
  - Score: 0.02
Solution: Evaluate units for semantic segmentation

Unit 1

Top activated images

Lamp, Intersection over Union (IoU) = 0.12

Network Dissection
Framework to interpret the deep visual representations

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
Result of 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

conv5 unit 200  mountain (object)  IoU=0.11

Histogram of object detectors: Detector:81/256, Unique Detector:40 (Units with IoU>0.04)
Are the emerging concepts real?


• “No distinction between individual high level units and random linear combinations of high level unit“
• “It suggests that it is the space, rather than the individual units, that contains the semantic information in network”
Are the emerging concepts real?
Are the emerging concepts real?
Are the emerging concepts real?

Random combination of units

Do concepts associate with individual units or the whole feature space?
Are the emerging concepts real?

AlexNet on Places365

AlexNet on ImageNet
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
3x3 conv, 256, pool/2
fc, 4096
fc, 4096
fc, 1000

GoogLeNet

ResNet

Datasets

IMAGENET

places

THE SCENE RECOGNITION DATABASE
Interpretable Units in Different Architectures

Number of Unique Detectors

- object
- part
- scene
- material
- texture
- color

AlexNet-random
AlexNet-ImageNet
AlexNet-Places365
Interpretable Units in Different Architectures

Number of Unique Detectors

- object
- part
- scene
- material
- texture
- color

AlexNet-random
AlexNet-ImageNet
AlexNet-Places365
VGG-ImageNet
GoogLeNet-ImageNet
GoogLeNet-Places365
VGG-Places365
Interpretable Units in Different Architectures

Number of Unique Detectors

- object
- part
- scene
- material
- texture
- color

AlexNet-random
AlexNet-ImageNet
AlexNet-Places365
VGG-ImageNet
GoogLeNet-ImageNet
GoogLeNet-Places365
VGG-Places365
ResNet152-ImageNet
ResNet152-Places365
Object units built from Places

AlexNet Detector: 81  Unique Detector: 40

ResNet Detector: 774  Unique Detector: 84

Object units built from ImageNet

AlexNet Detector: 49  Unique Detector: 21

ResNet Detector: 858  Unique Detector: 75
Interpretable Units over Layers

AlexNet on Places365

Number of unique detectors

conv1  conv2  conv3  conv4  conv5

object  part  scene  material  texture  color
Interpretable Units over Layers
Interpretable Units over Layers
CNNs Trained from Self-supervised Learning

Training CNN without image labels.

Context prediction, ICCV’15

Solving puzzle, ECCV’16

Colorization, ECCV’16 and CVPR’17

Audio prediction, ECCV’16
Comparison of Supervisions

All use AlexNet architecture
Comparison of Supervisions

All use AlexNet architecture

Number of unique detectors

AlexNet-Places365
AlexNet-ImageNet
tracking
object-centric
audio
moving
colorization
puzzle
crosschannel
egomotion
context
frameorder
AlexNet-random

object
part
scene
material
texture
color
Interpretable Units in Self-supervised Networks

Predict audio from video frames. ECCV’16 Owens et al.
Interpretable Units in Self-supervised Networks

Colorize grey images
ECCV’16. Zhang et al.

conv5 unit 15: banded (texture) \( \text{IoU}=0.13 \)

conv5 unit 159: tree (object) \( \text{IoU}=0.039 \)

conv5 unit 210: head (part) \( \text{IoU}=0.038 \)
Emergence of Interpretable Units during Training

Number of unique detectors

Accuracy on validation set

Training iteration 1

Training iteration 1
Individual Unit during Training

Unit 23 at conv5 layer
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
Explainable Deep Features

Activations from CNN as generic visual feature
Deep features as generic visual descriptor
Explaining the Output

Softmax class weights

Activation of units as object detectors

Explaining the Output

- Class Activation Maps (CAM) for the top 5 predictions: palace, dome, church, altar, monastery

Explaining the Output by Unit Interpretations

Walking the dog

Top activated units
Explaining the Output by Unit Interpretations

Top activated units
Explaining the Output by Unit Interpretations

Correct label: Gardening
Cutting vegetables

Top activated units
Conclusion

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

Living room
Kitchen
Coast
Theater
...

Network Dissection

Interpretability Report

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