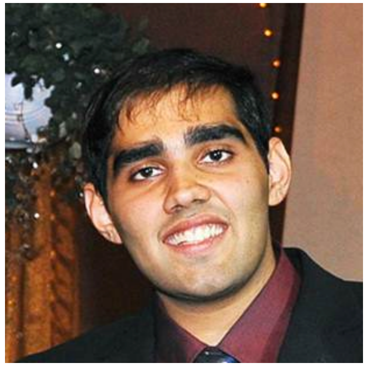
Interpreting Deep Visual Representations

Bolei Zhou

MIT

David Bau

Aditya Khosla

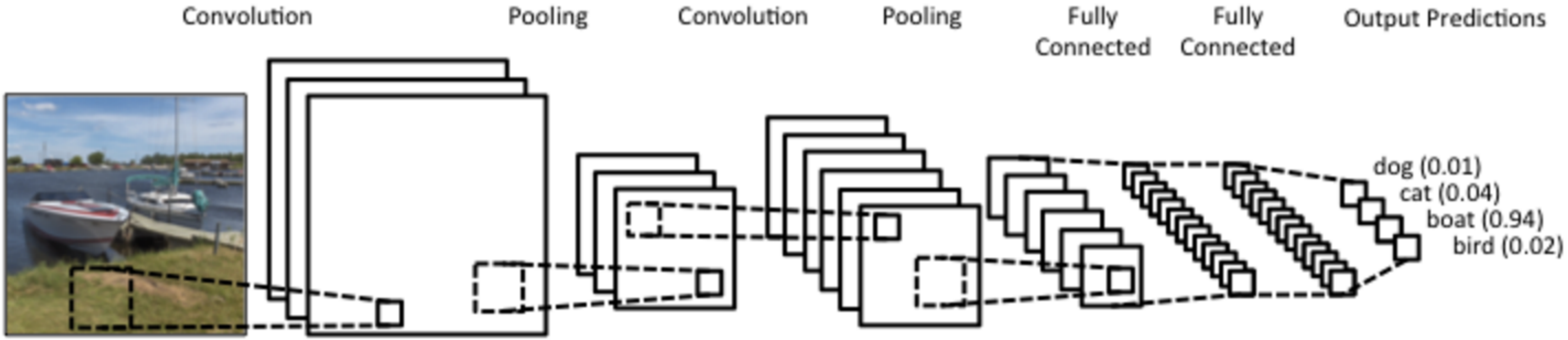


Aude Oliva

Antonio Torralba

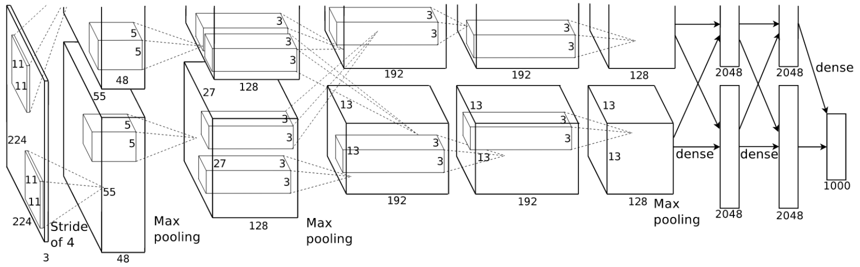
Background

Convolutional Neural Network (ConvNet)

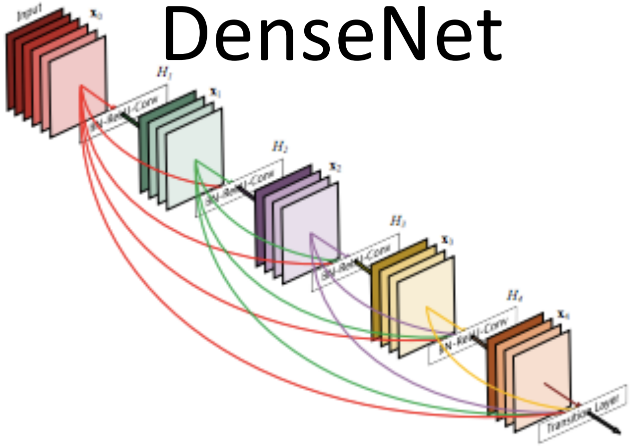


Many networks

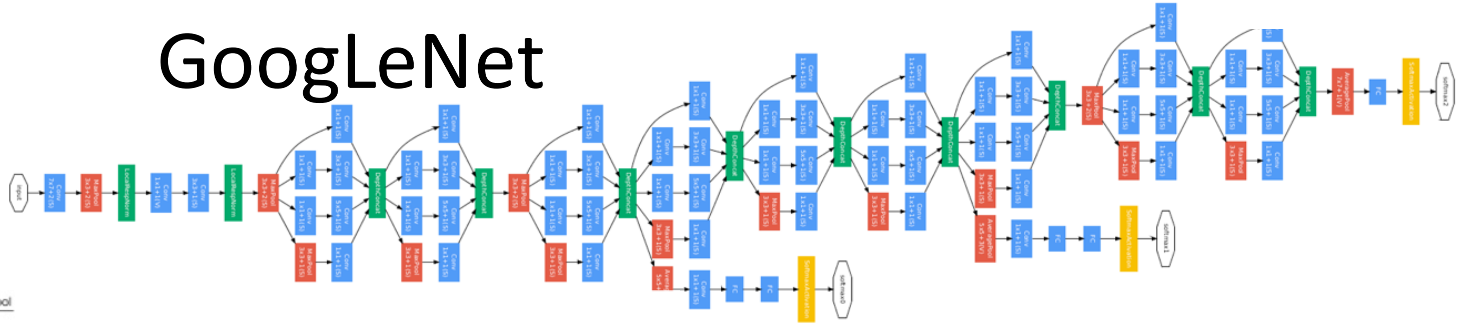
AlexNet

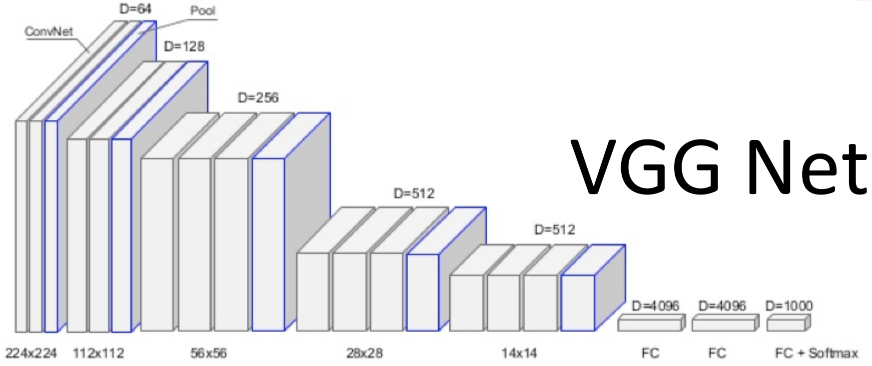


DenseNet



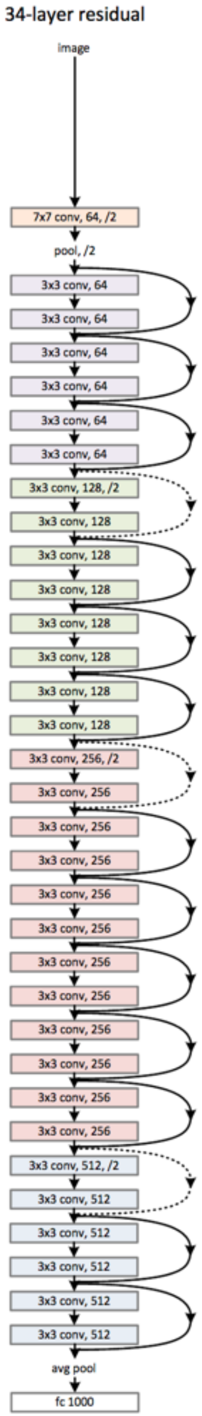
GoogLeNet



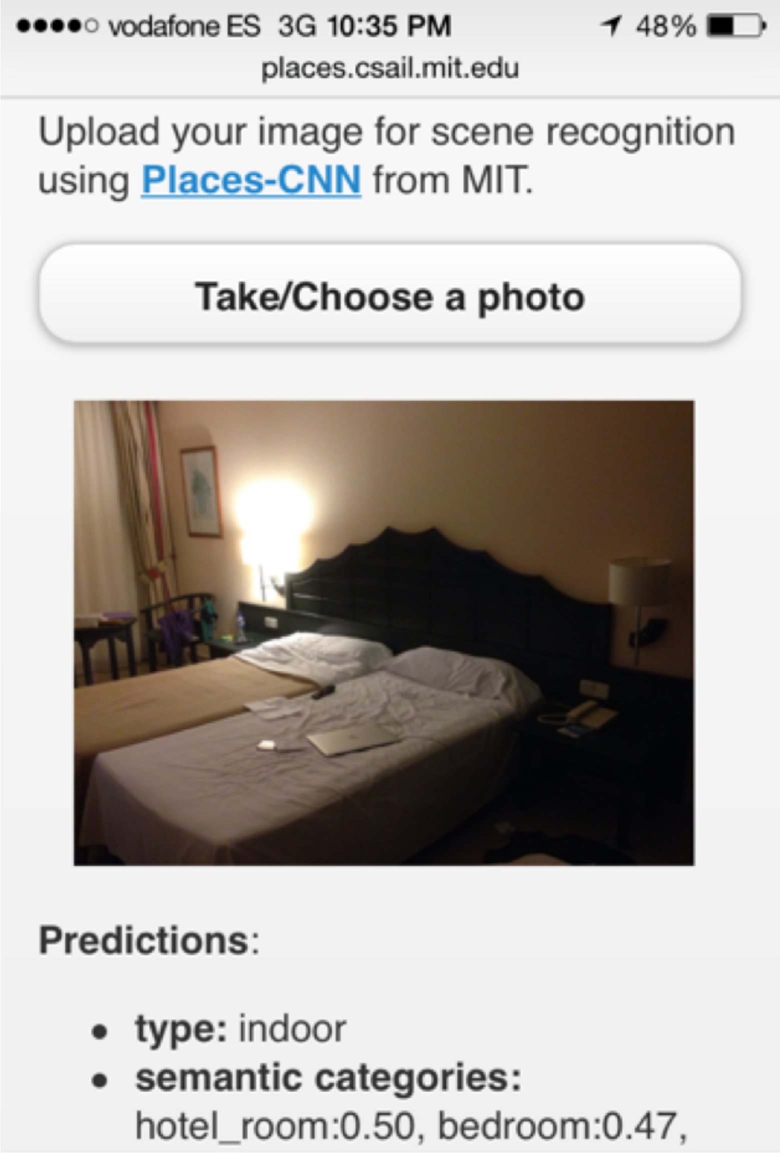


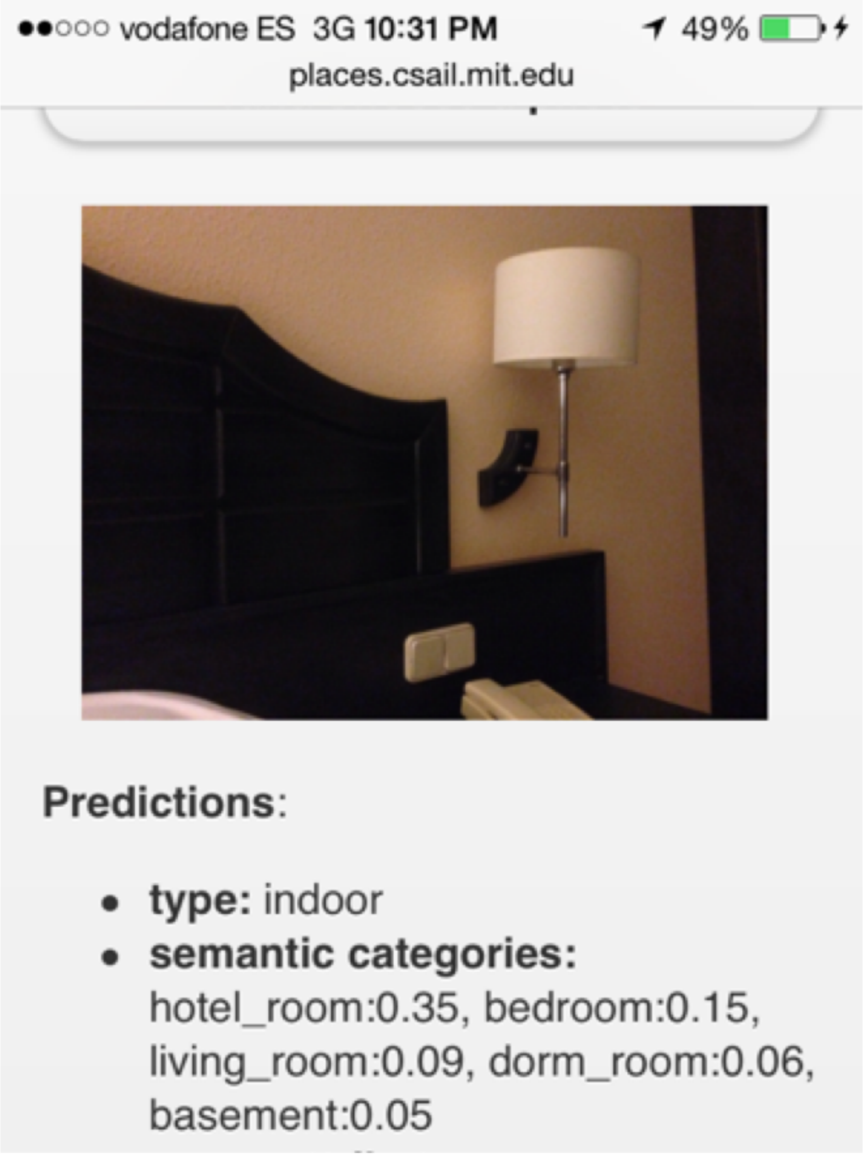
VGG Net

ResNet



Why works so well





When it fails, why is it?

Output: cutting vegetables.
Correct label: gardening

Output: washing dishes.
Correct label: brushing

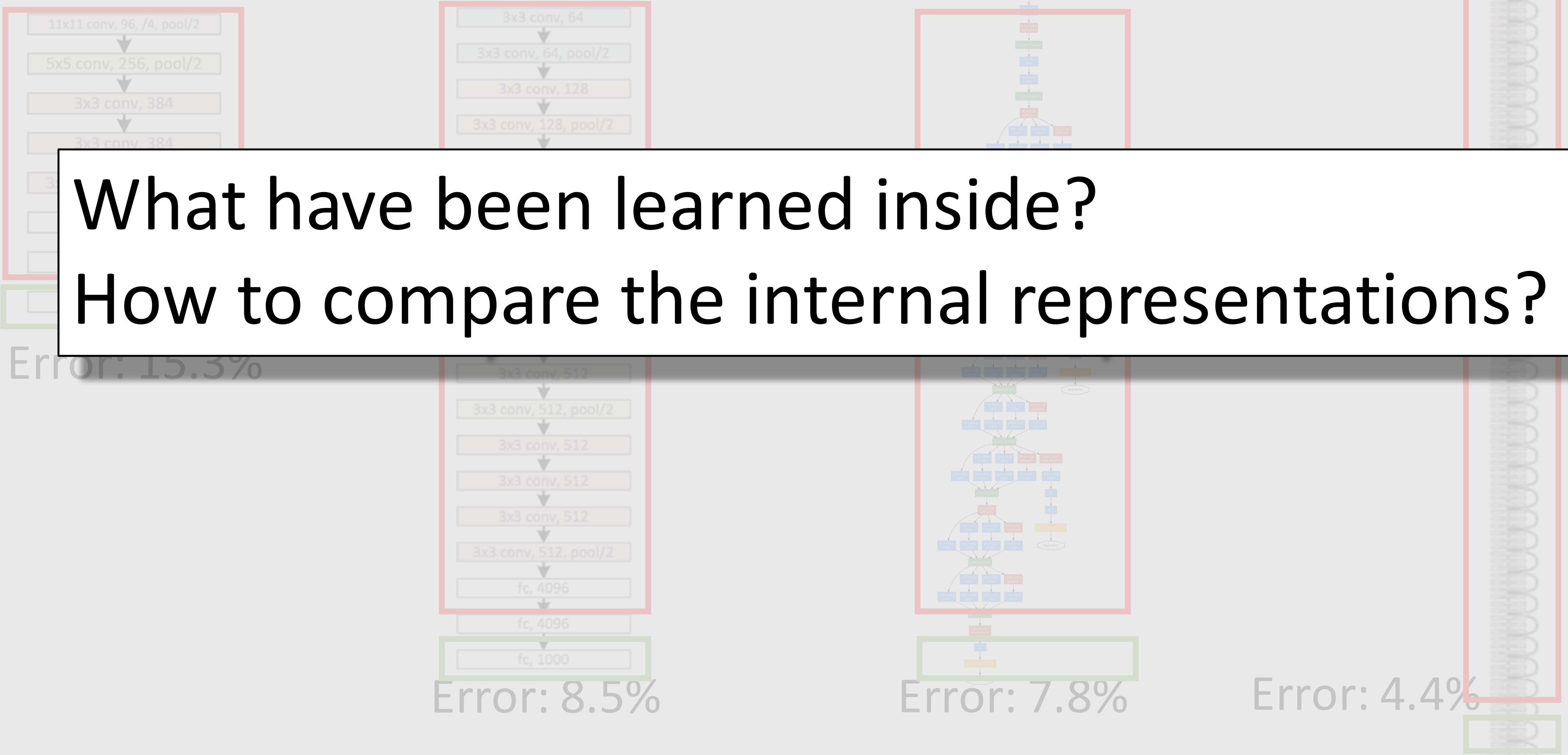
Deep ConvNet for Visual Recognition

2012: AlexNet
5 conv. layers

2014: VGG
16 conv. layers

2015: GoogLeNet
22 conv. layers

2016: ResNet
>100 conv. layers



Error: 15.3%

Error: 8.5%

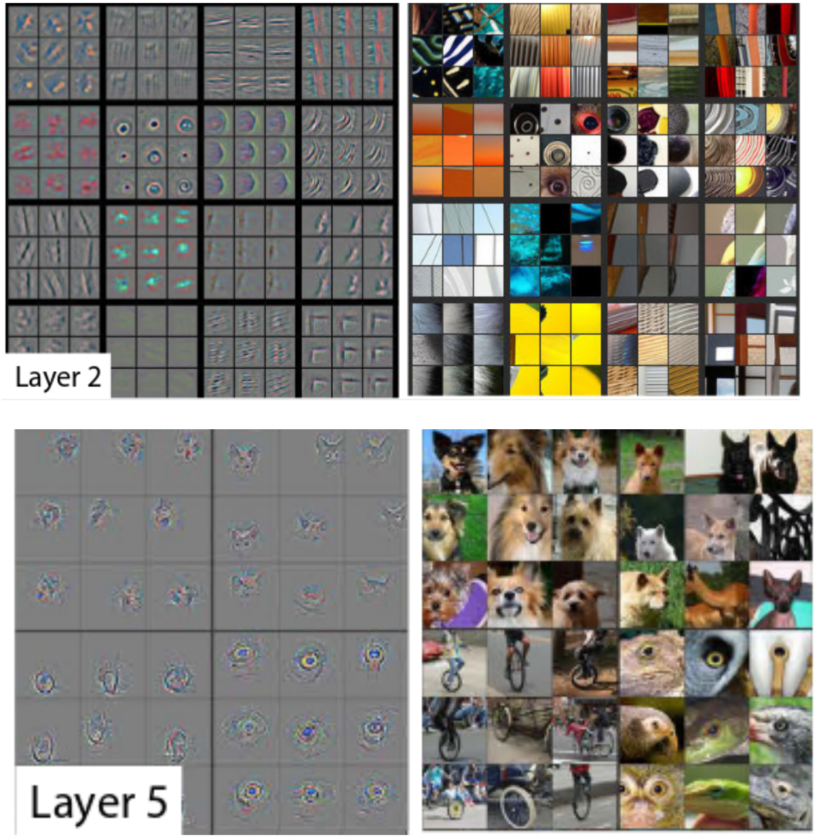
Error: 7.8%

Error: 4.4%

What have been learned inside?
How to compare the internal representations?

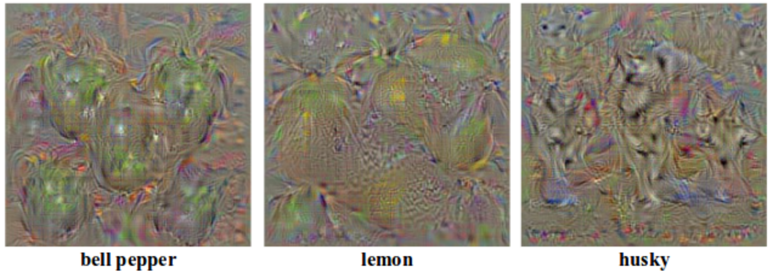
Work on Network Visualization

Deconvolution



Zeiler et al., ECCV 2014.

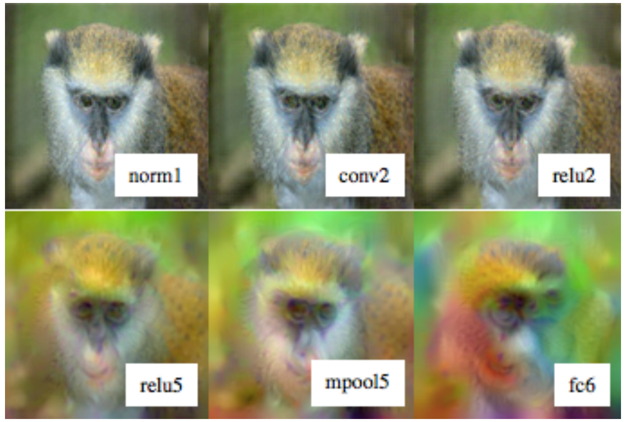
Back-propagation



Simonyan et al., ICLR 2015

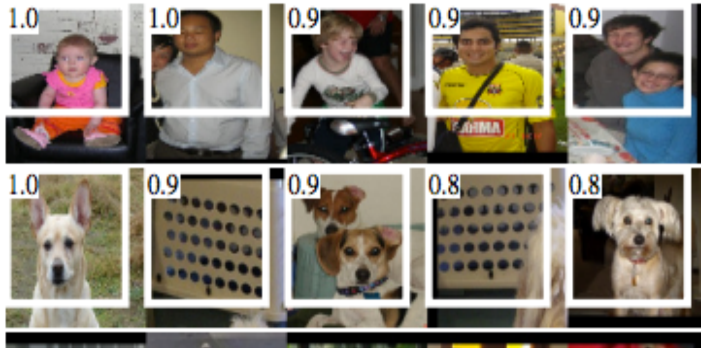
Inceptionism. Google Blog. June 2015

Feature inversion



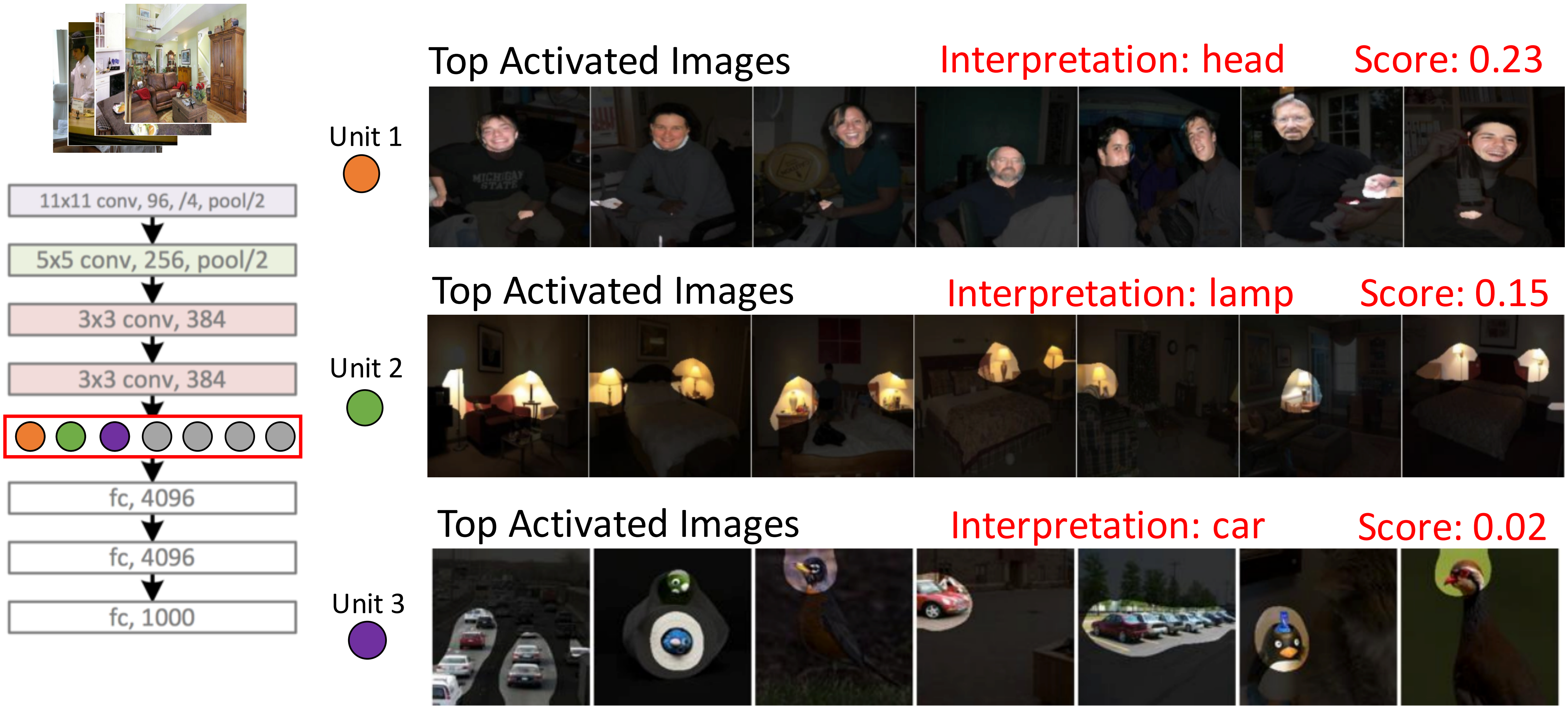
Mahendran et al, CVPR 2015

Top activated images



Girshick et al., CVPR 2014

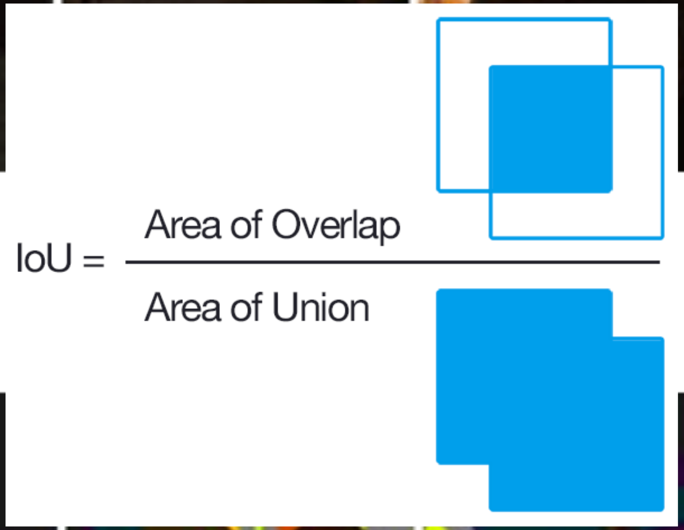
Going From Visualization to Interpretation



Solution: Evaluate units for semantic segmentation

Unit 1 Top activated images

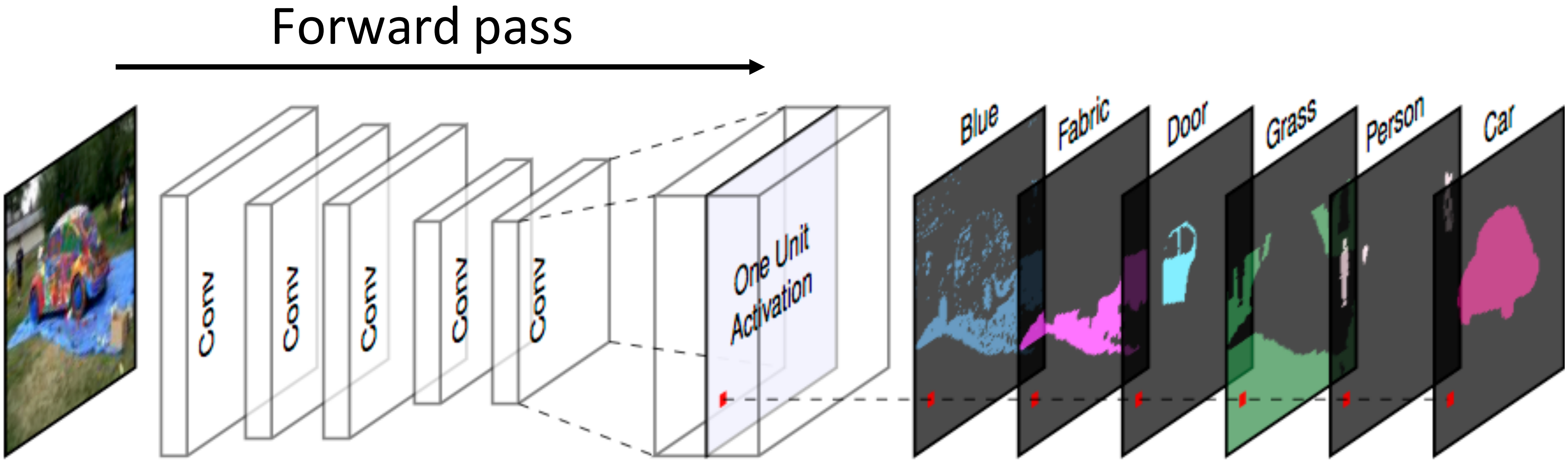
Lamp, Intersection over Union (IoU)= 0.12



Bau*, Zhou*, Khosla, Oliva, Torralba. Network Dissection: quantifying interpretability of deep visual representations. CVPR’17

Network Dissection

Framework to interpret the deep visual representations



Bau*, Zhou*, Khosla, Oliva, Torralba. Network Dissection: quantifying interpretability of deep visual representations. CVPR'17

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

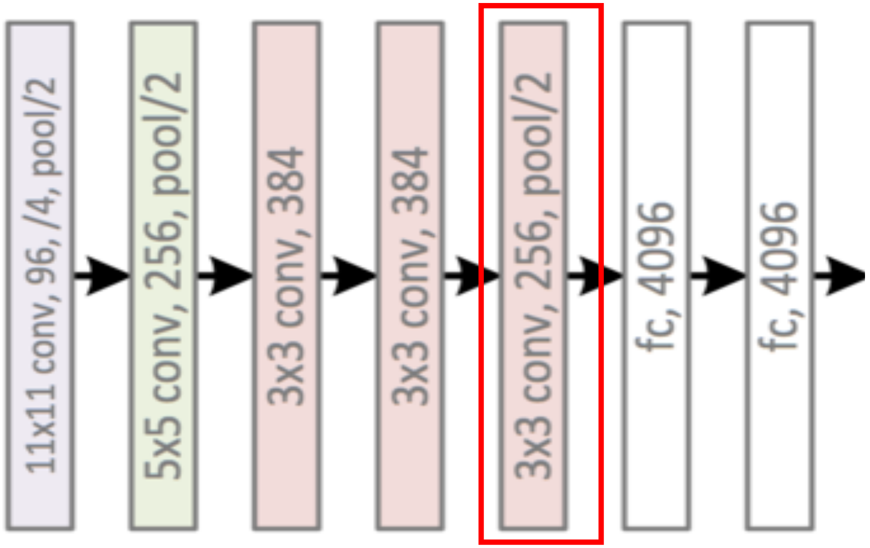
Total = 63,305 images

1,197 visual concepts

Result of AlexNet trained on places

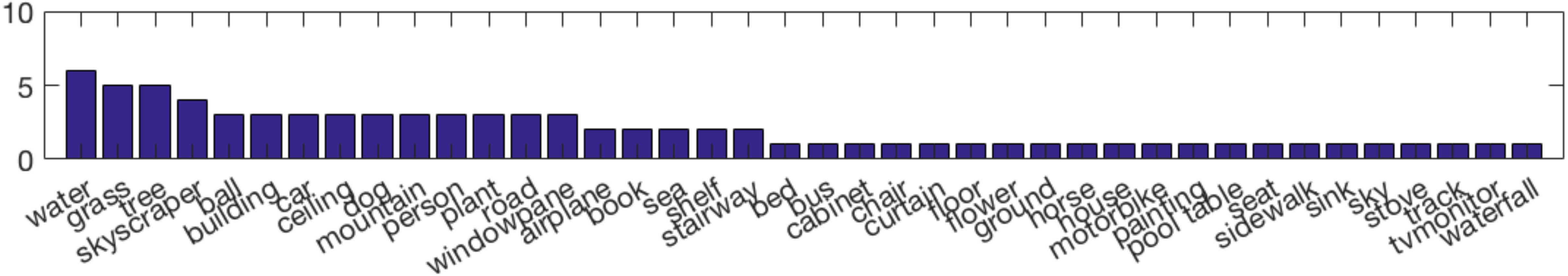
THE SCENE RECOGNITION DATABASE

conv5, 256 units



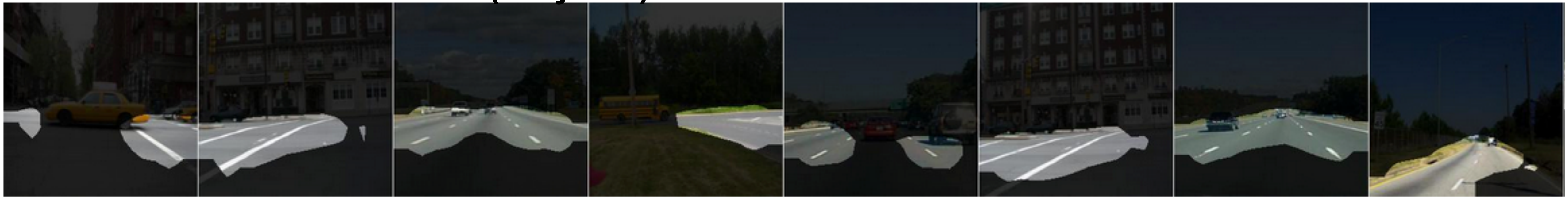
Living room
Kitchen
Coast
…
365 categories

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

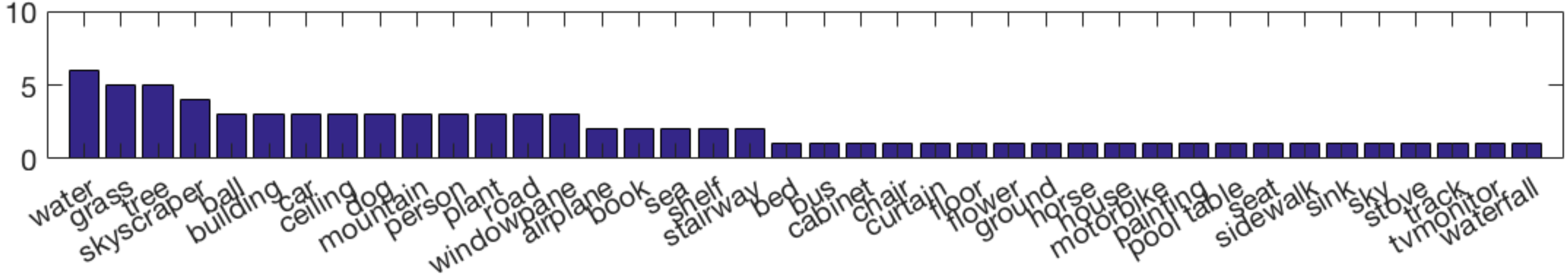


conv5 unit 79 car (object) IoU=0.13

conv5 unit 107 road (object) IoU=0.15

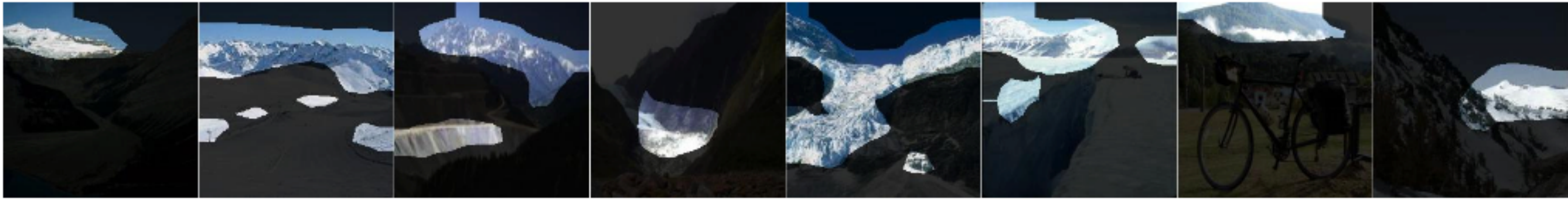


Histogram of object detectors: Detector:81/256, Unique Detector:40 (Units with IoU>0.04)

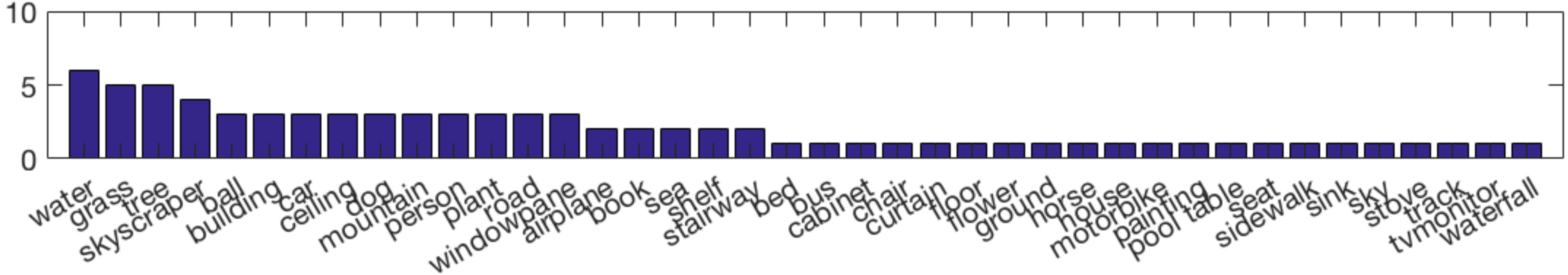


conv5 unit 144 mountain (object) IoU=0.13

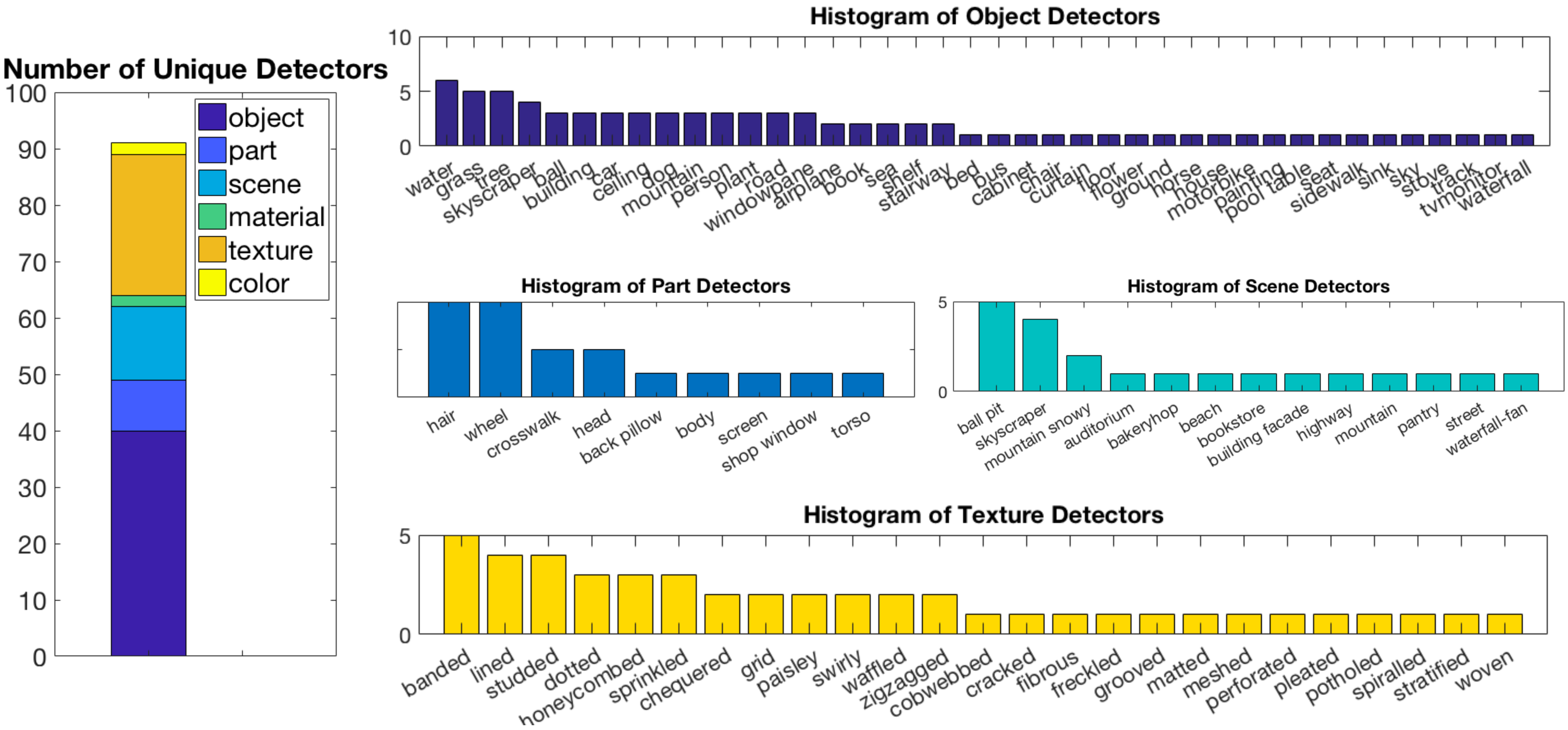
conv5 unit 200 mountain (object) IoU=0.11



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



Dissection Report



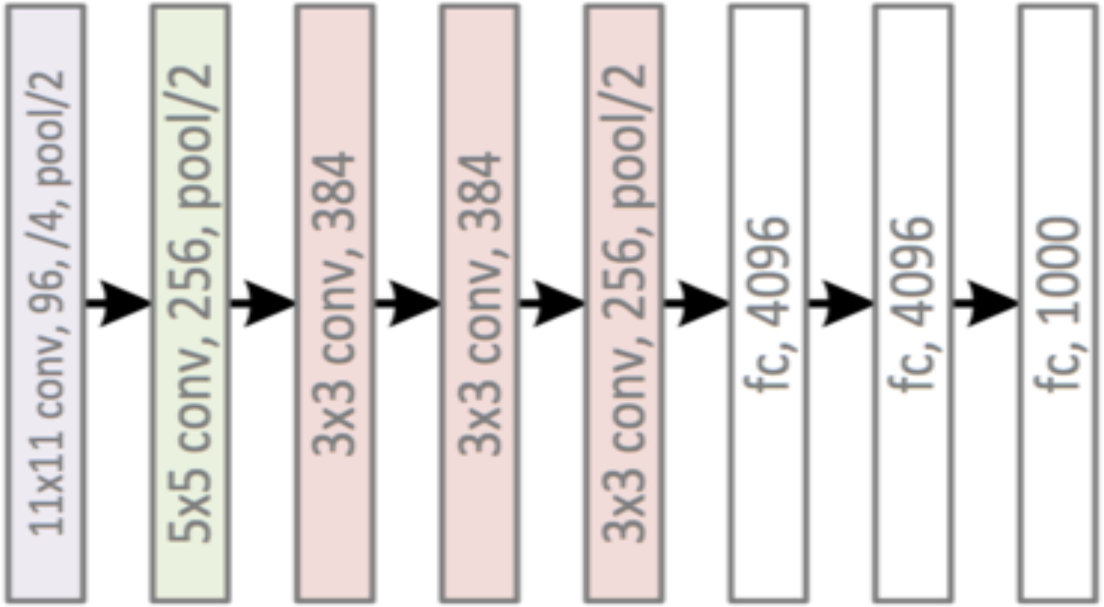
Are the emerging concepts real?

Szegedy et al. Intriguing properties of neural networks. arXiv.2014

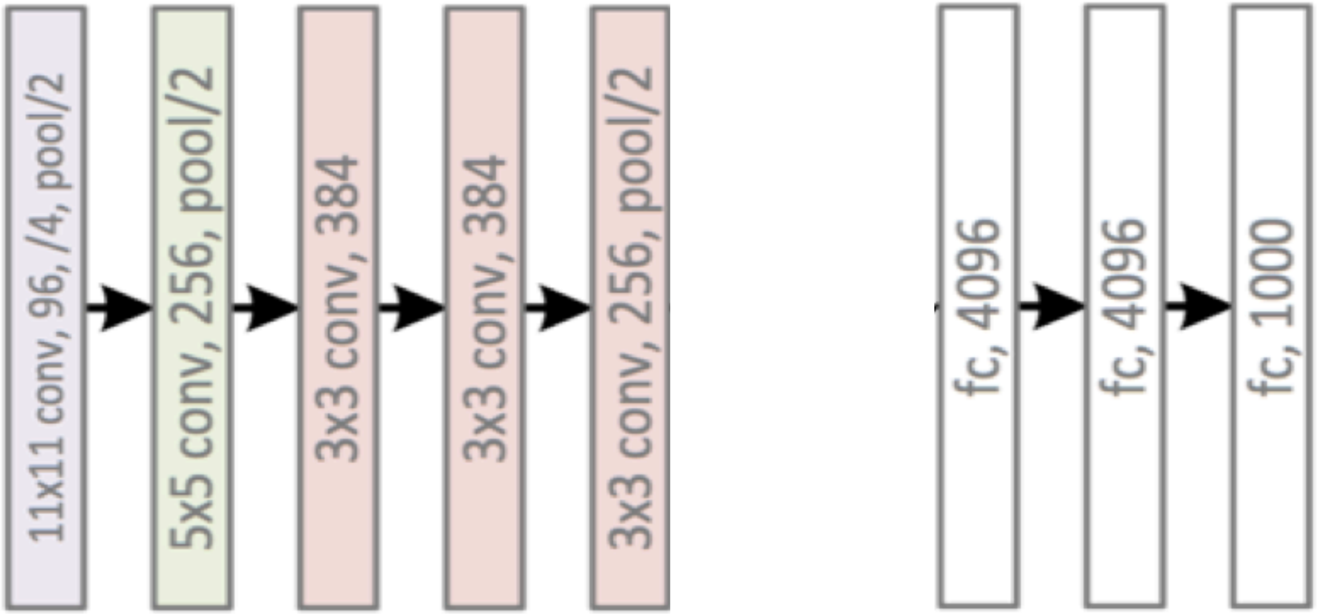
- "No distinction between individual high level units and random linear combinations of high level unit"
- "It suggests that it is the space, rather than the individual units, that contains the semantic information in network"

Single Neuron Random Projection

Are the emerging concepts real?

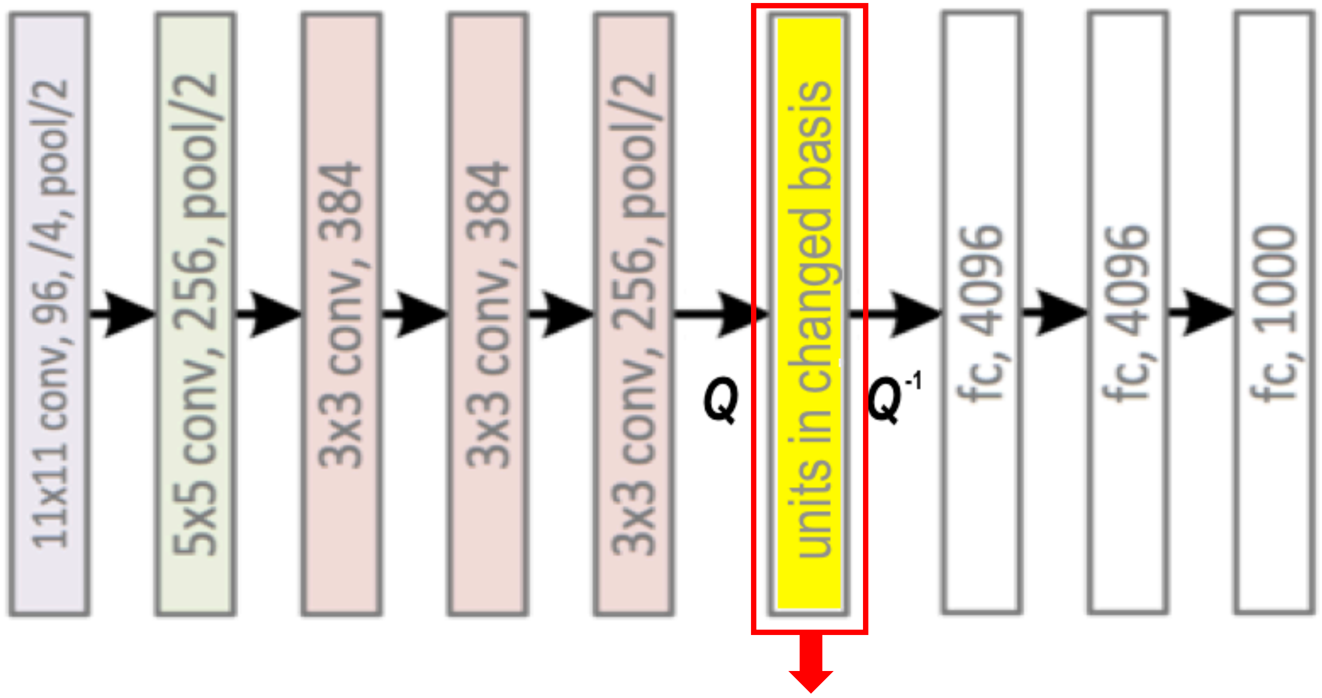


Are the emerging concepts real?



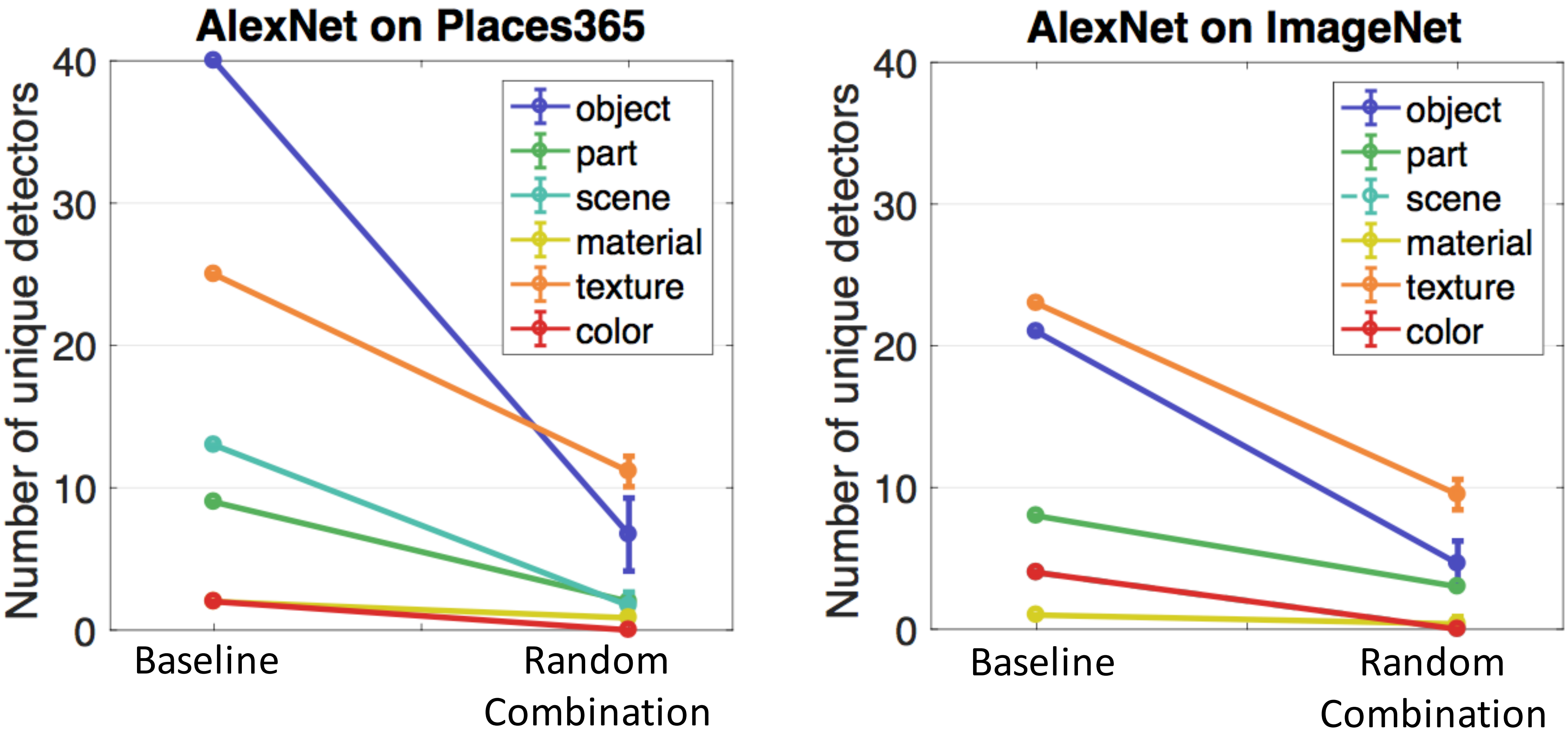
Are the emerging concepts real?

Random combination of units

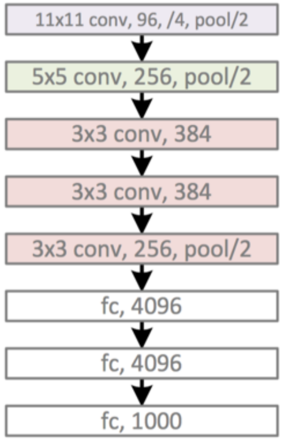


Do concepts associate with individual units or the whole feature space?

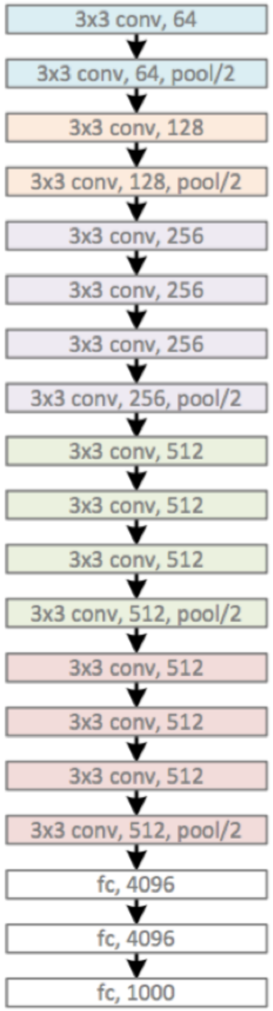
Are the emerging concepts real?



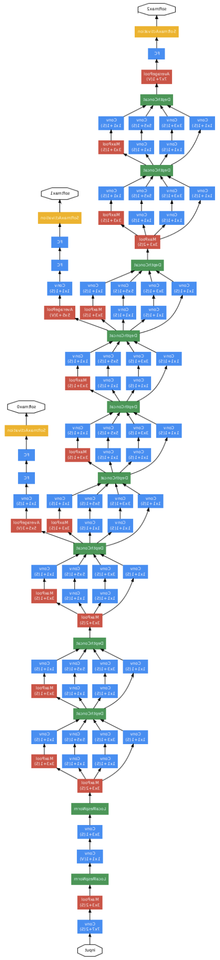
Architectures



AlexNet



VGG



GoogLeNet

ResNet

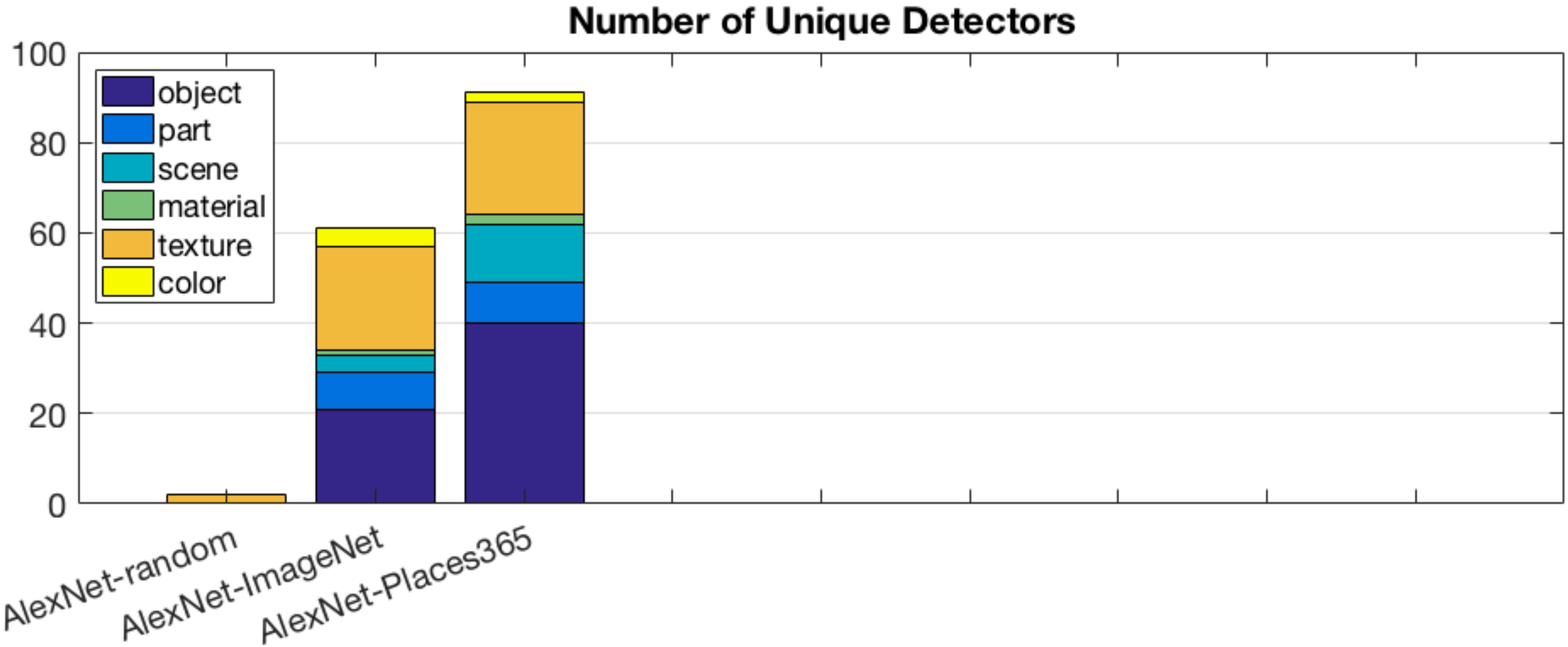
Datasets

IMAGENET

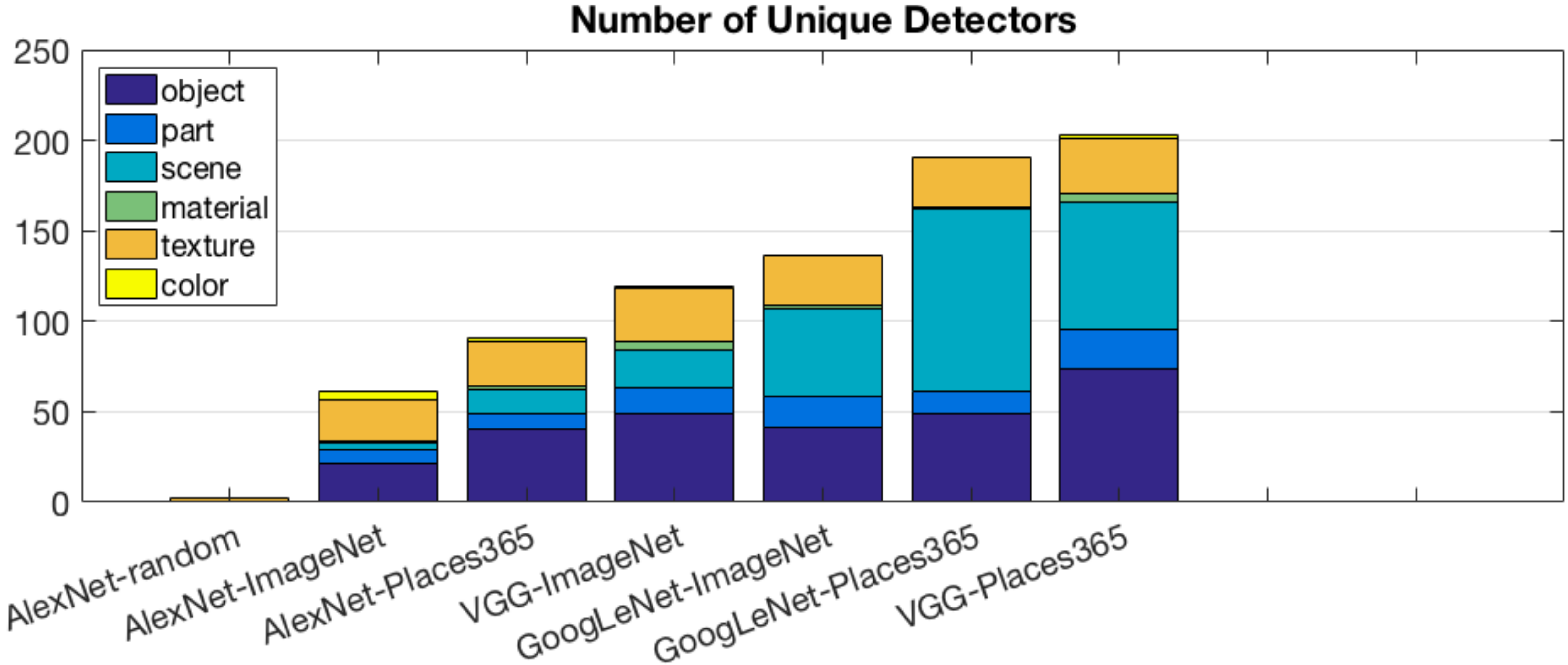
places

THE SCENE RECOGNITION DATABASE

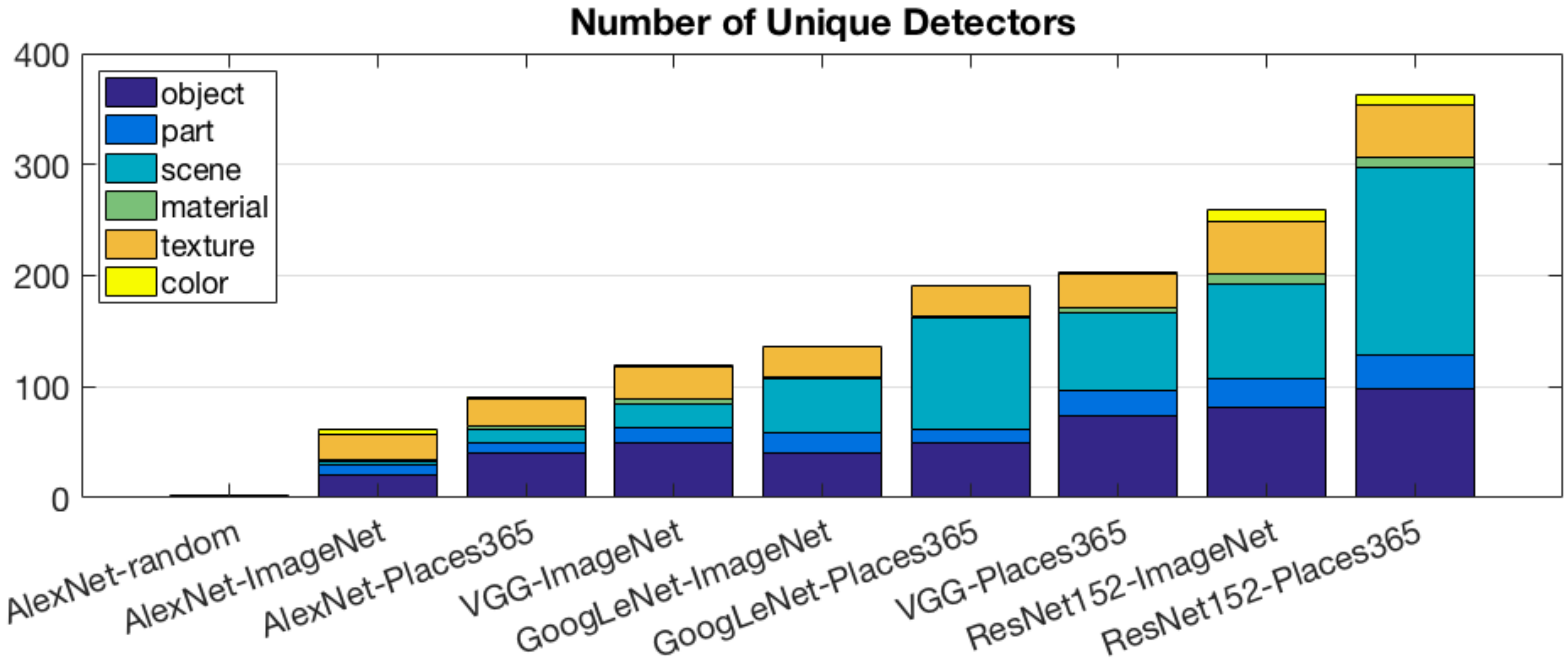
Interpretable Units in Different Architectures

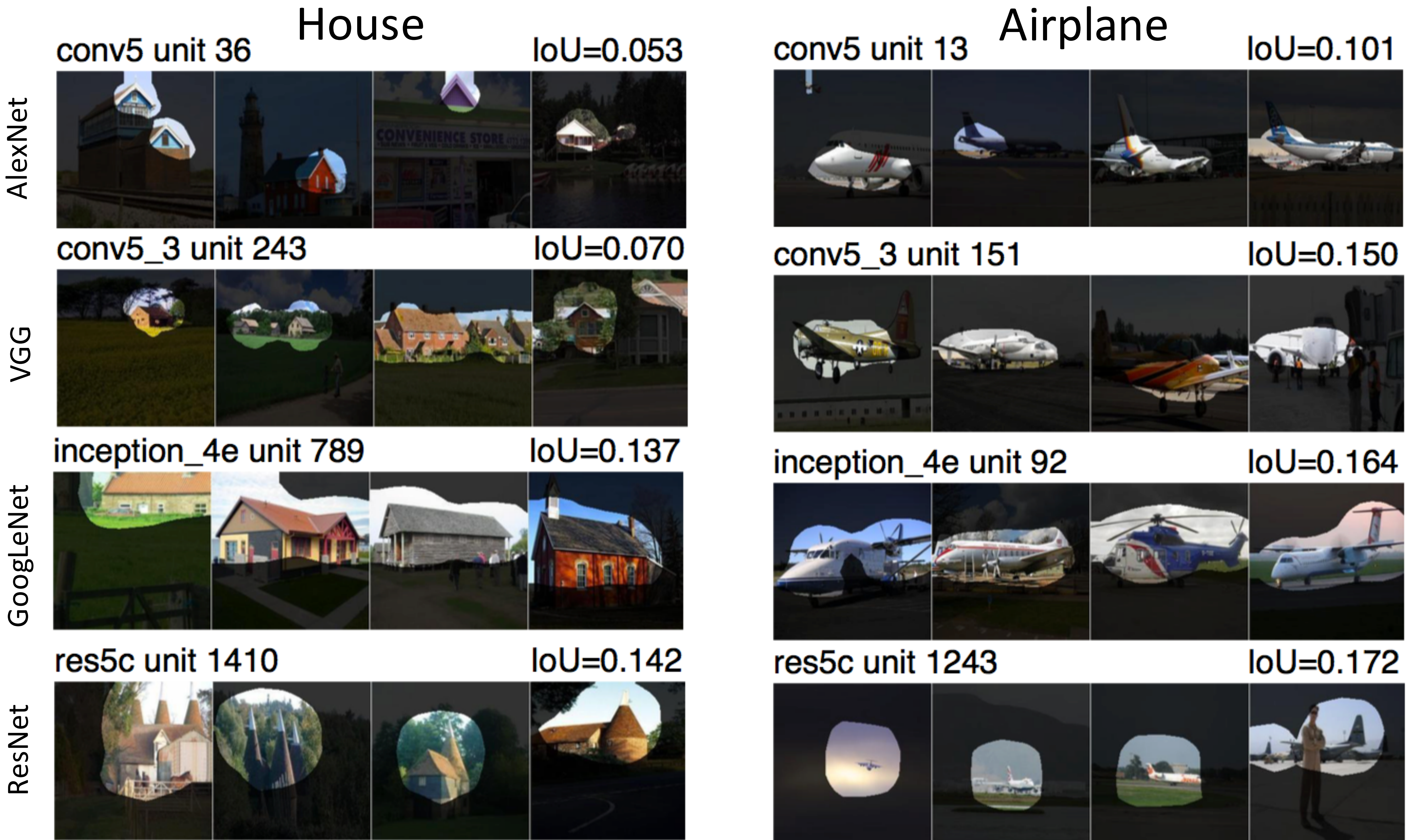


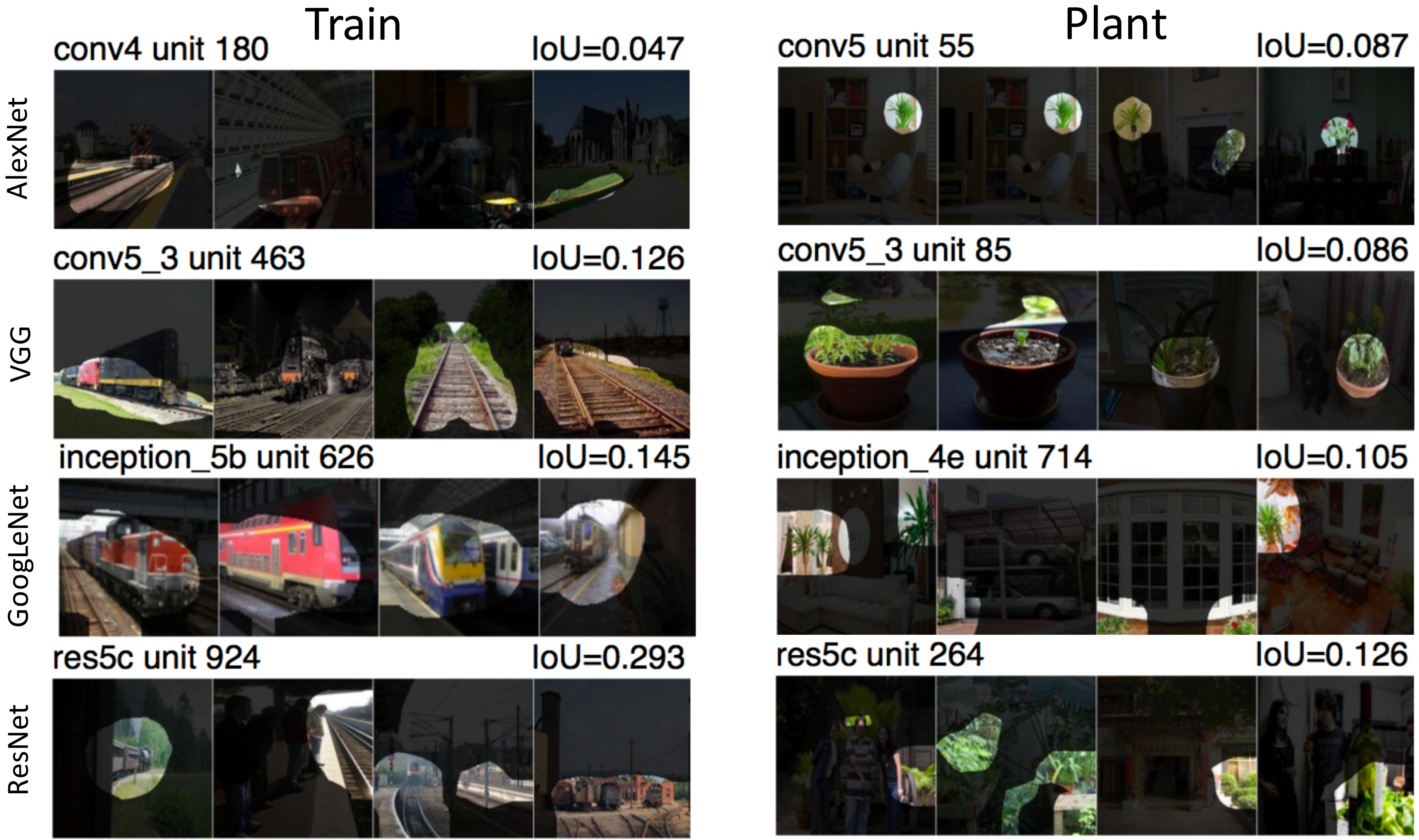
Interpretable Units in Different Architectures



Interpretable Units in Different Architectures

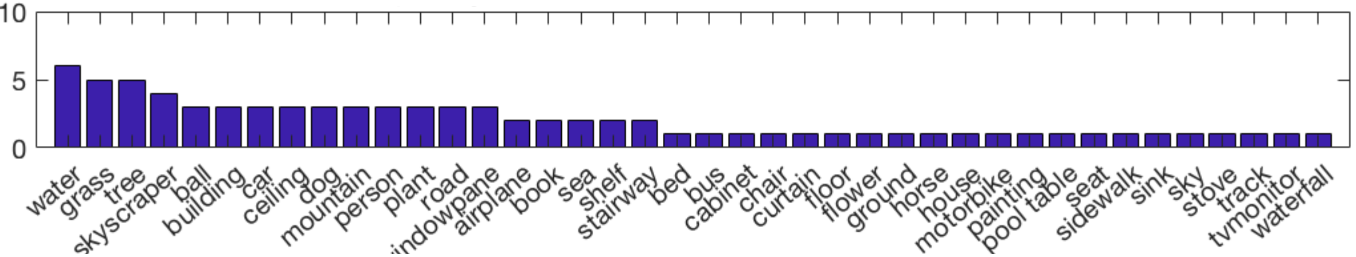



House
Airplane
AlexNet
VGG
GoogLeNet
ResNet
conv5 unit 36
IoU=0.053
conv5 unit 13
IoU=0.101
conv5_3 unit 243
IoU=0.070
conv5_3 unit 151
IoU=0.150
inception_4e unit 789
IoU=0.137
inception_4e unit 92
IoU=0.164
res5c unit 1410
IoU=0.142
res5c unit 1243
IoU=0.172
CONVENIENCE STORE



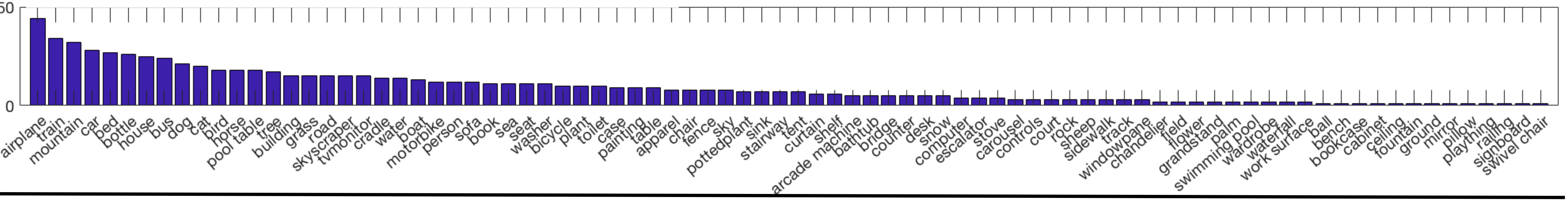
Train
Plant
AlexNet
VGG
GoogLeNet
ResNet
conv4 unit 180
IoU=0.047
conv5 unit 55
IoU=0.087
conv5_3 unit 463
IoU=0.126
conv5_3 unit 85
IoU=0.086
inception_5b unit 626
IoU=0.145
inception_4e unit 714
IoU=0.105
res5c unit 924
IoU=0.293
res5c unit 264
IoU=0.126

AlexNet Detector: 81 Unique Detector: 40

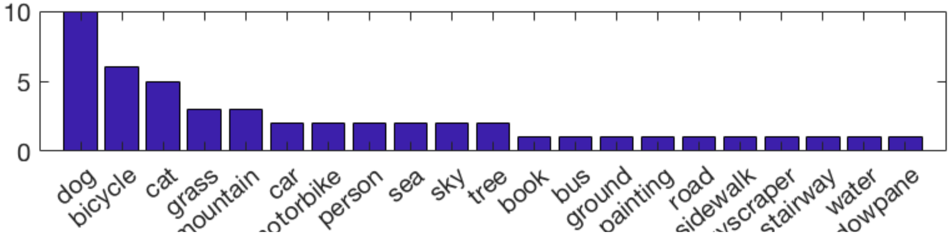


Object units built from Places

ResNet Detector: 774 Unique Detector: 84

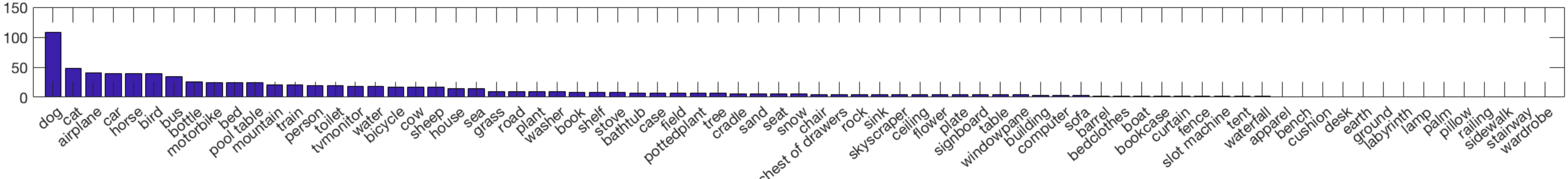


AlexNet Detector: 49 Unique Detector: 21

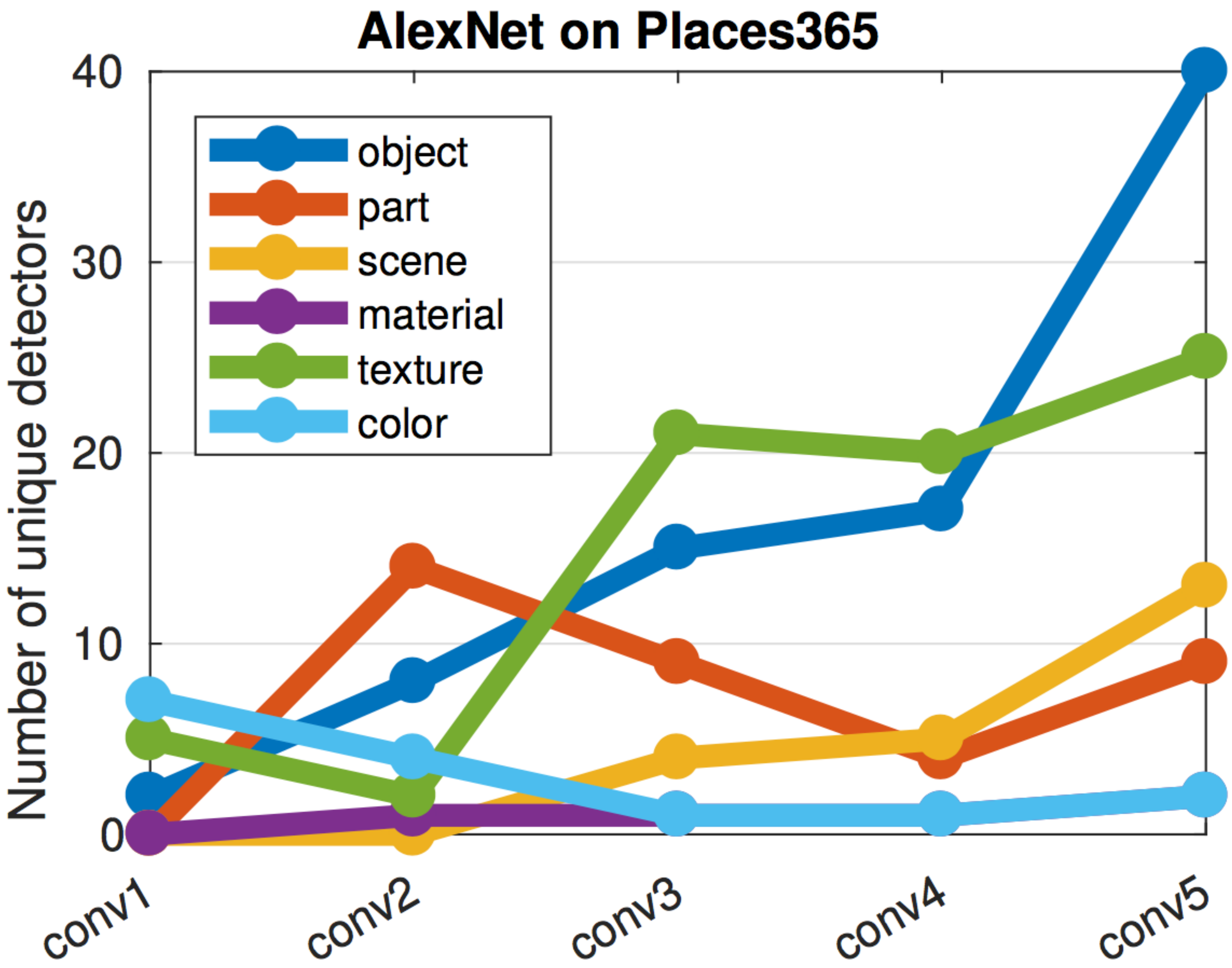


Object units built from ImageNet

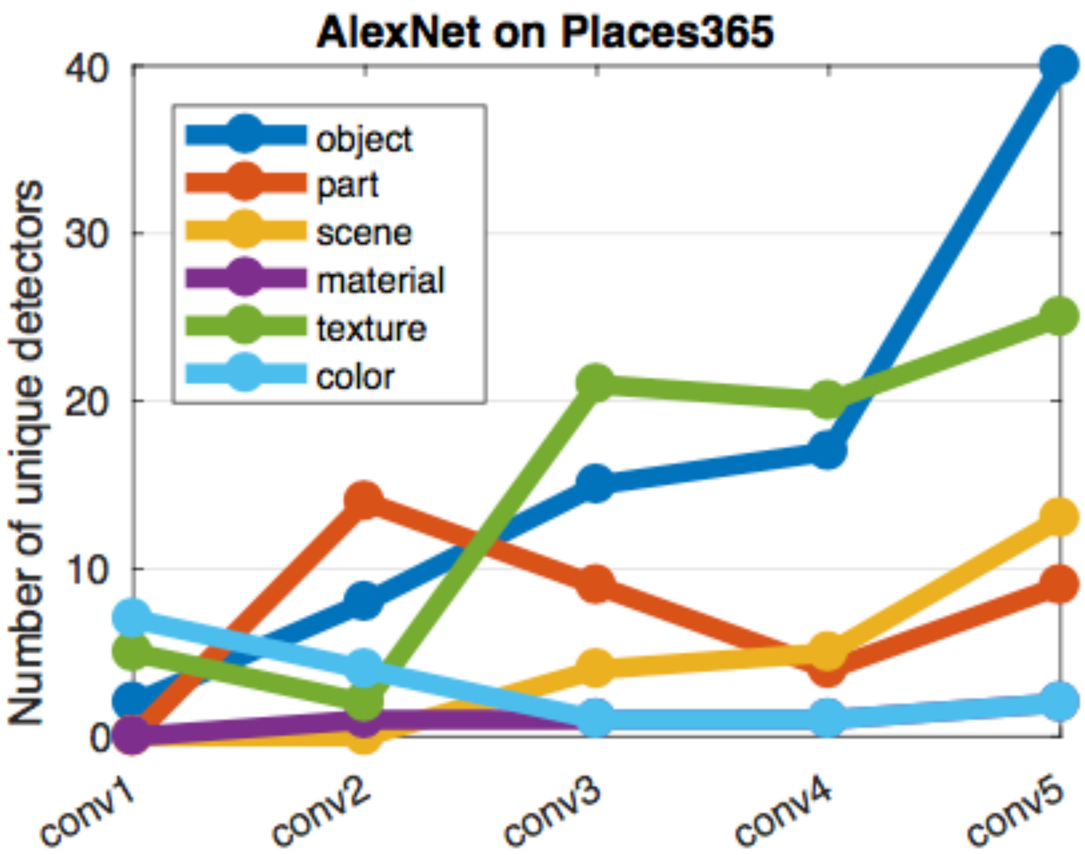
ResNet Detector: 858 Unique Detector: 75

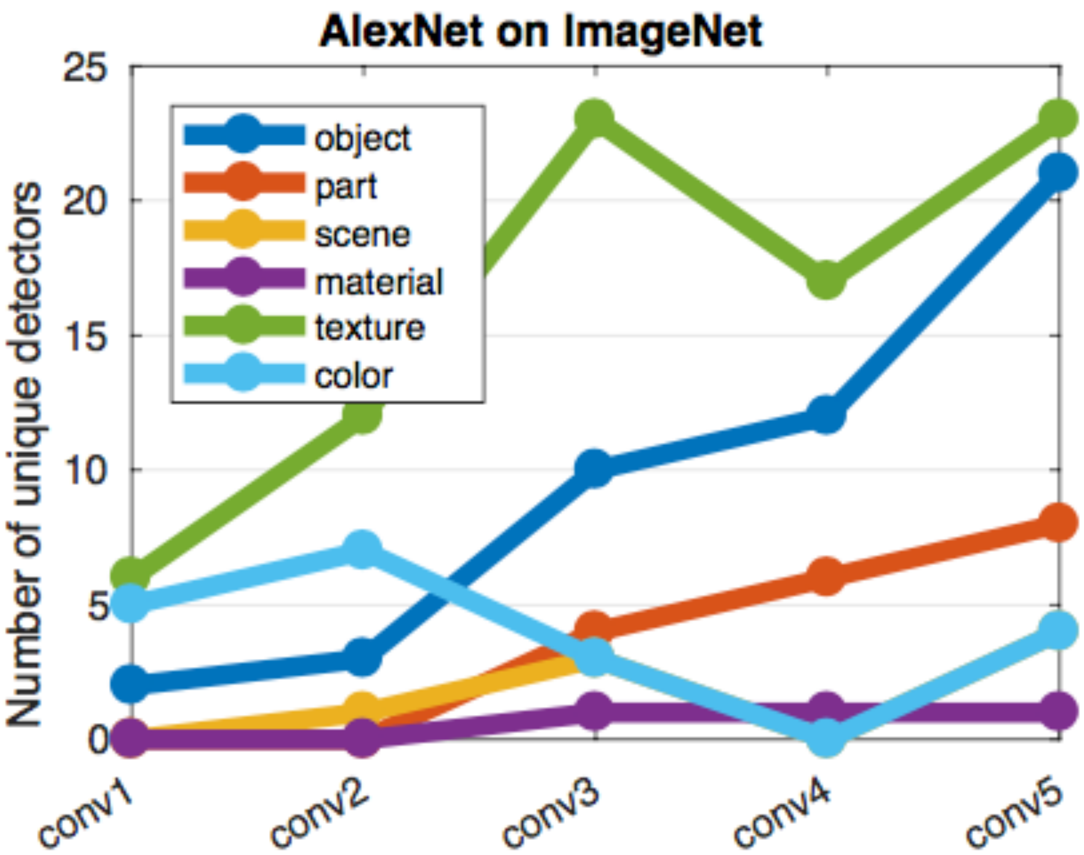


Interpretable Units over Layers

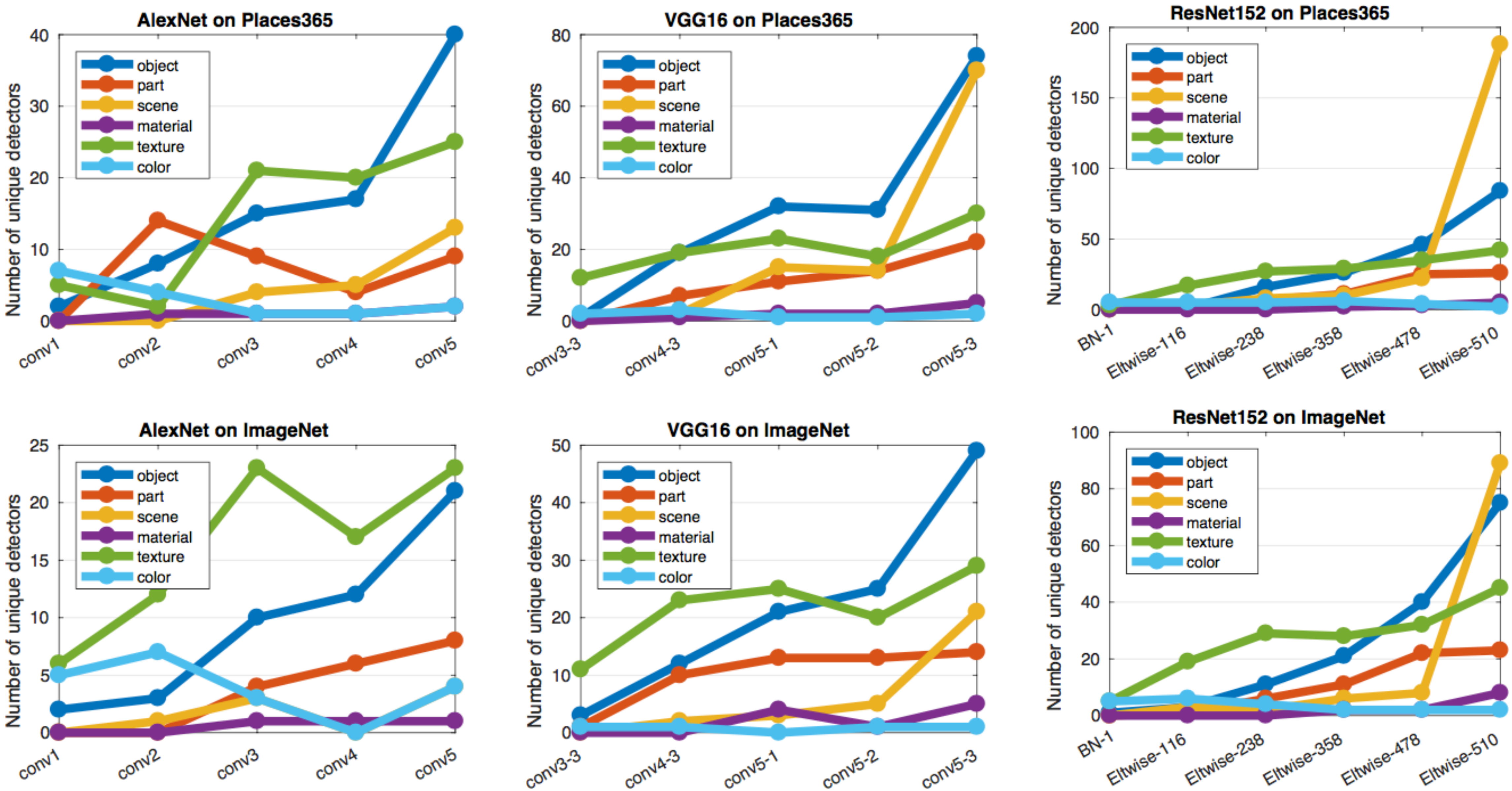


Interpretable Units over Layers





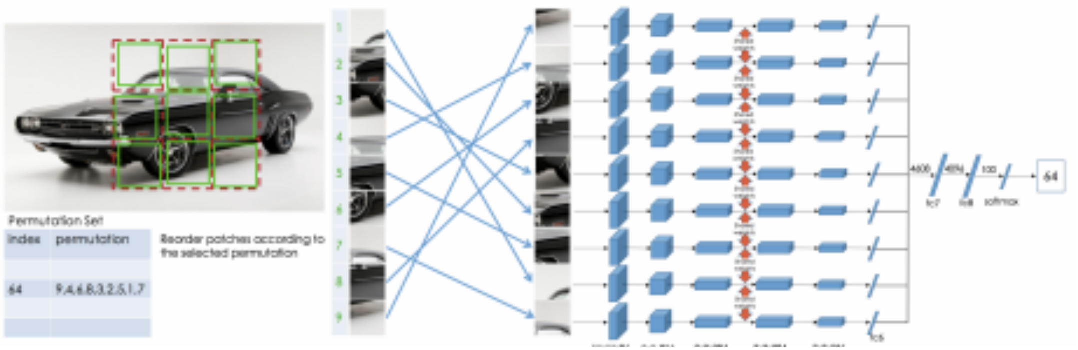
Interpretable Units over Layers



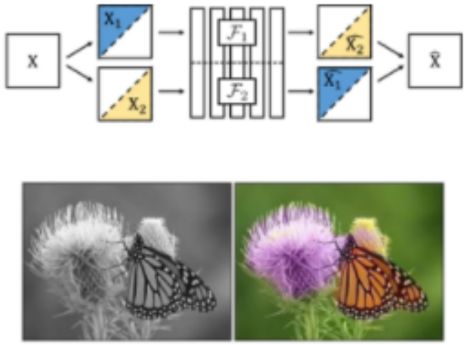
CNNs Trained from Self-supervised Learning

Training CNN without image labels.

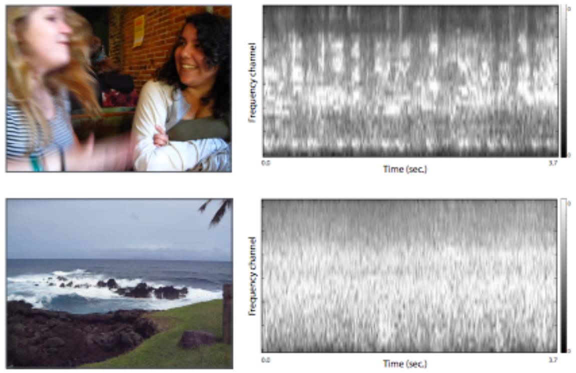
Context prediction, ICCV'15



Solving puzzle, ECCV'16



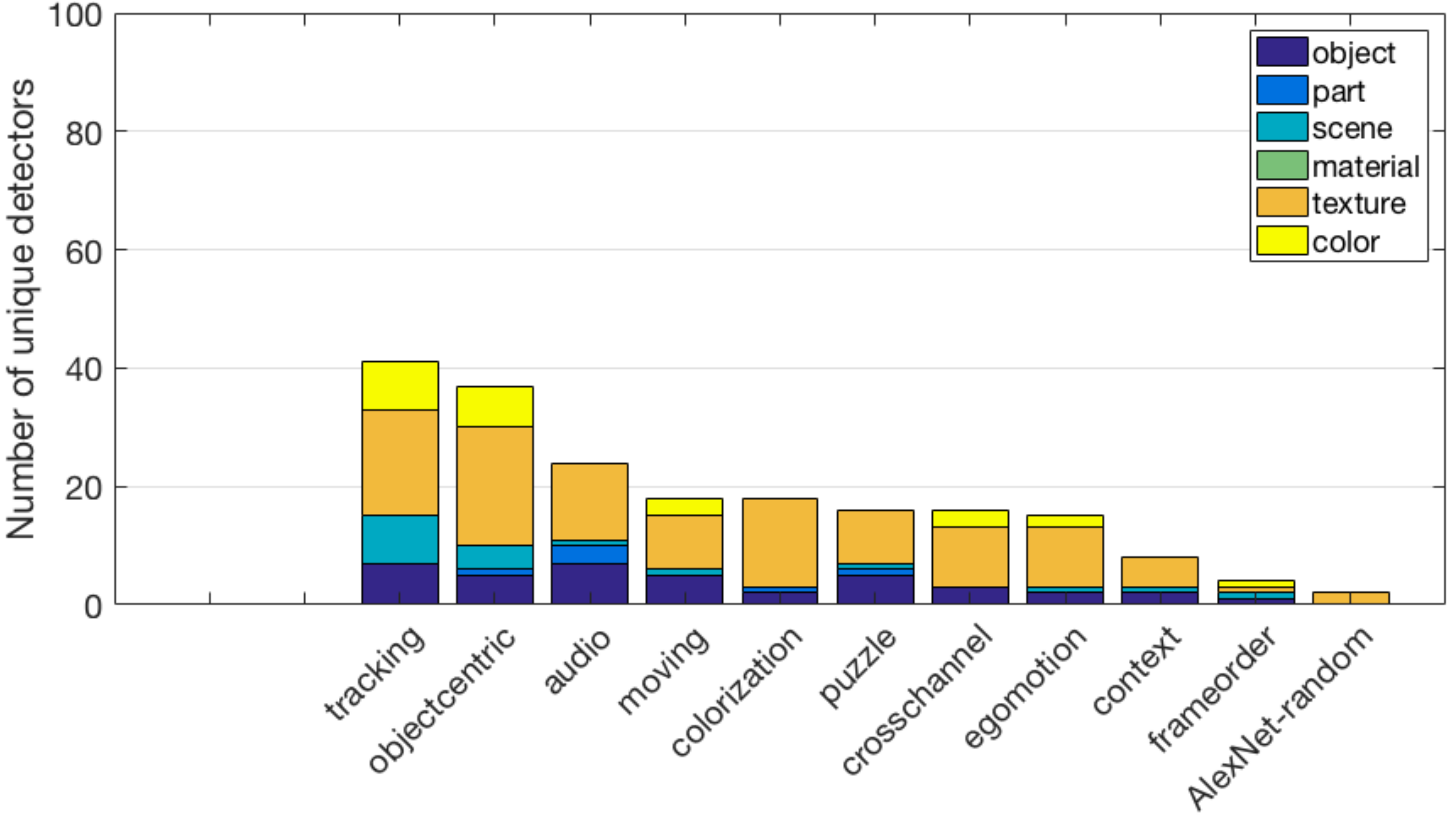
Colorization, ECCV'16 and CVPR'17



(a) Video frame (b) Cochleagram

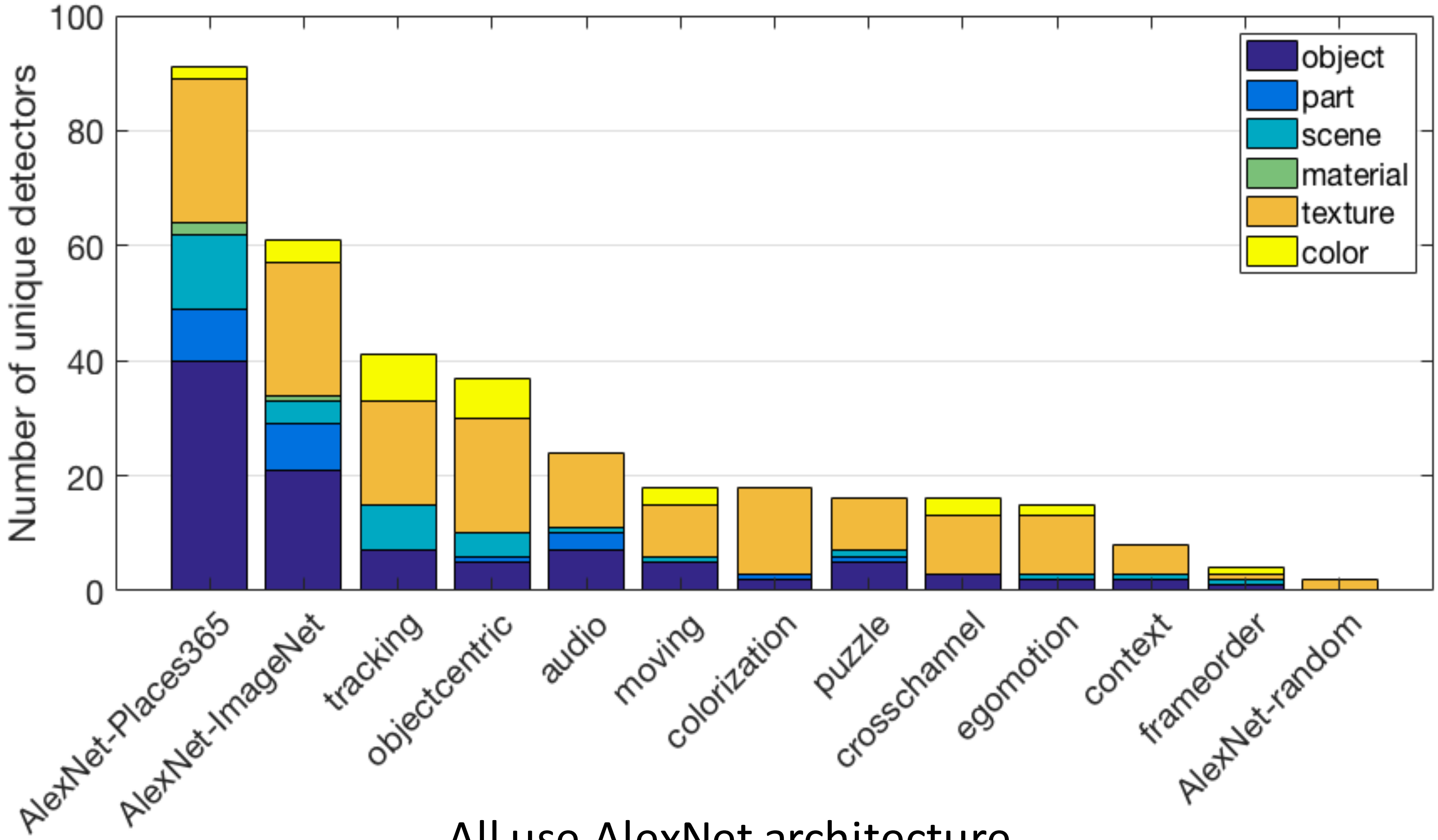
Audio prediction, ECCV'16

Comparison of Supervisions



All use AlexNet architecture

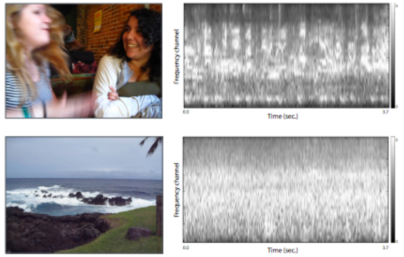
Comparison of Supervisions

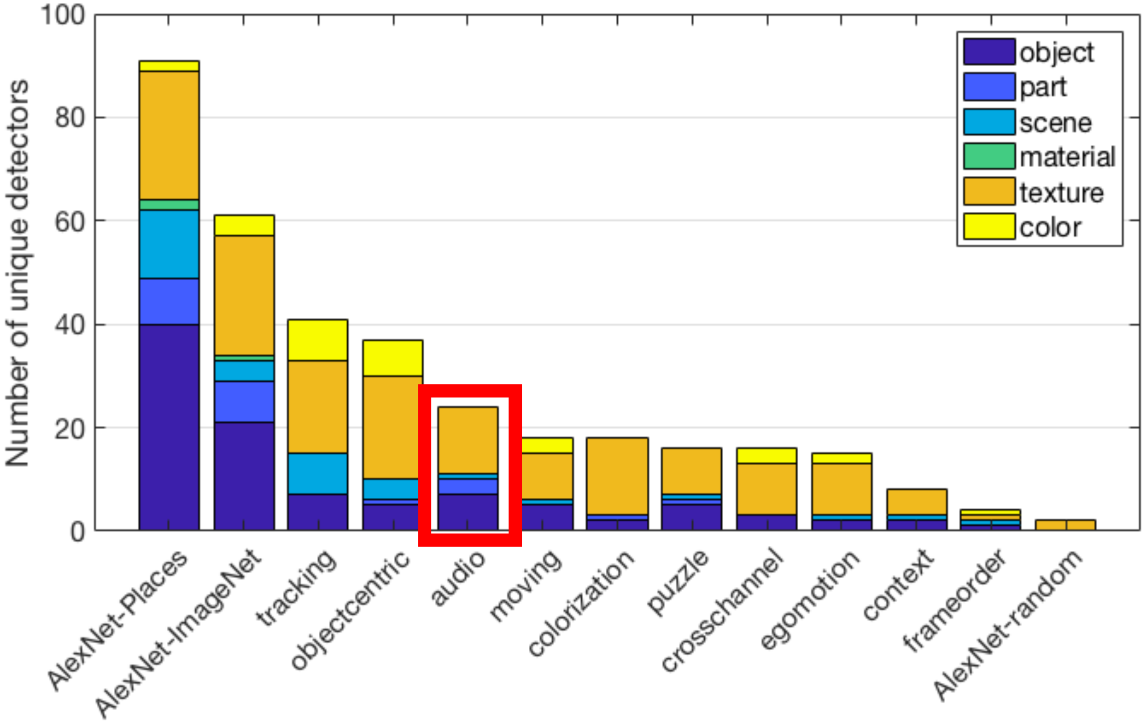


All use AlexNet architecture

Interpretable Units in Self-supervised Networks

Predict audio from video frames. ECCV'16 Owens et al.

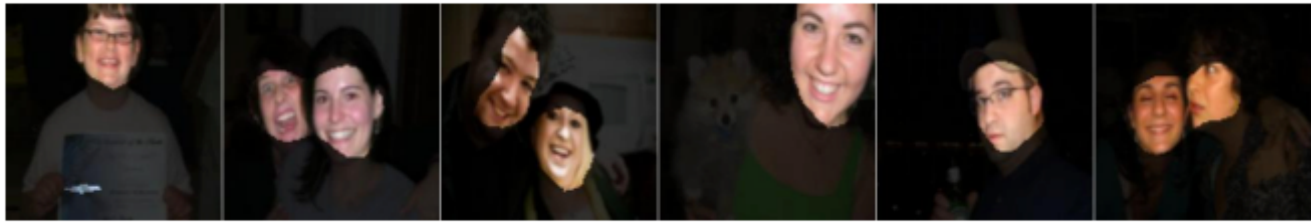




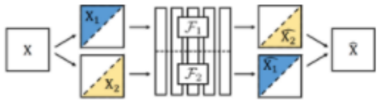
conv5 unit 205: car (object)   IoU=0.063

conv5 unit 124: creek (scene)   IoU=0.031

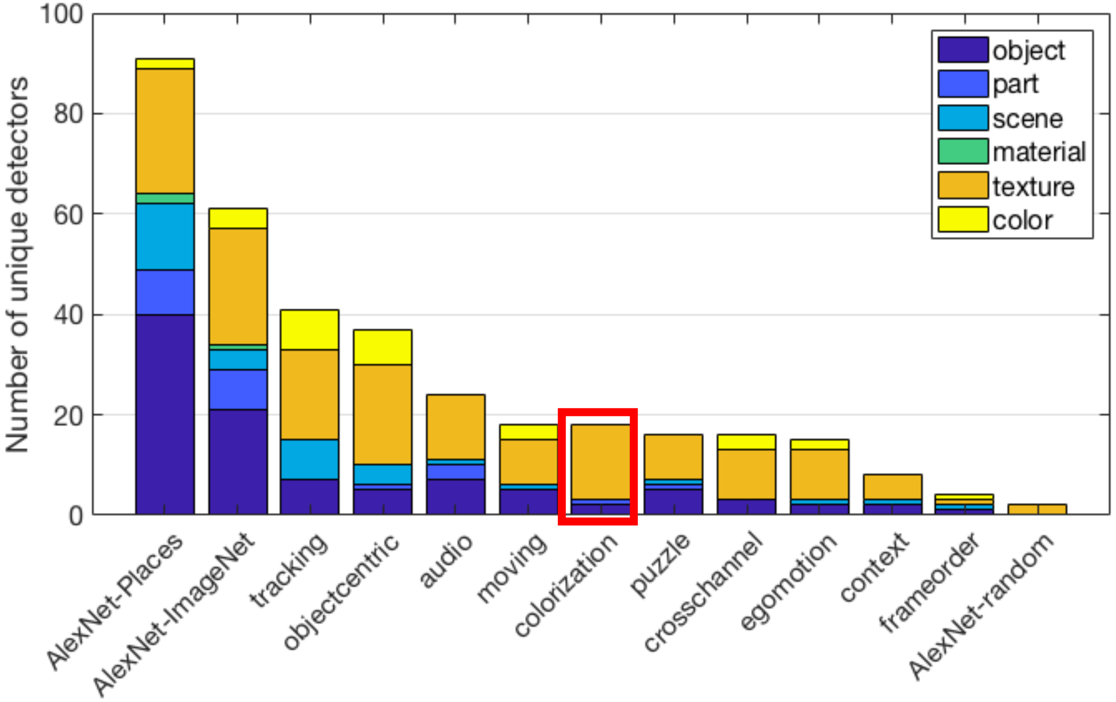
conv5 unit 51: head (part)   IoU=0.061



Interpretable Units in Self-supervised Networks



Colorize grey images
ECCV'16. Zhang et al.

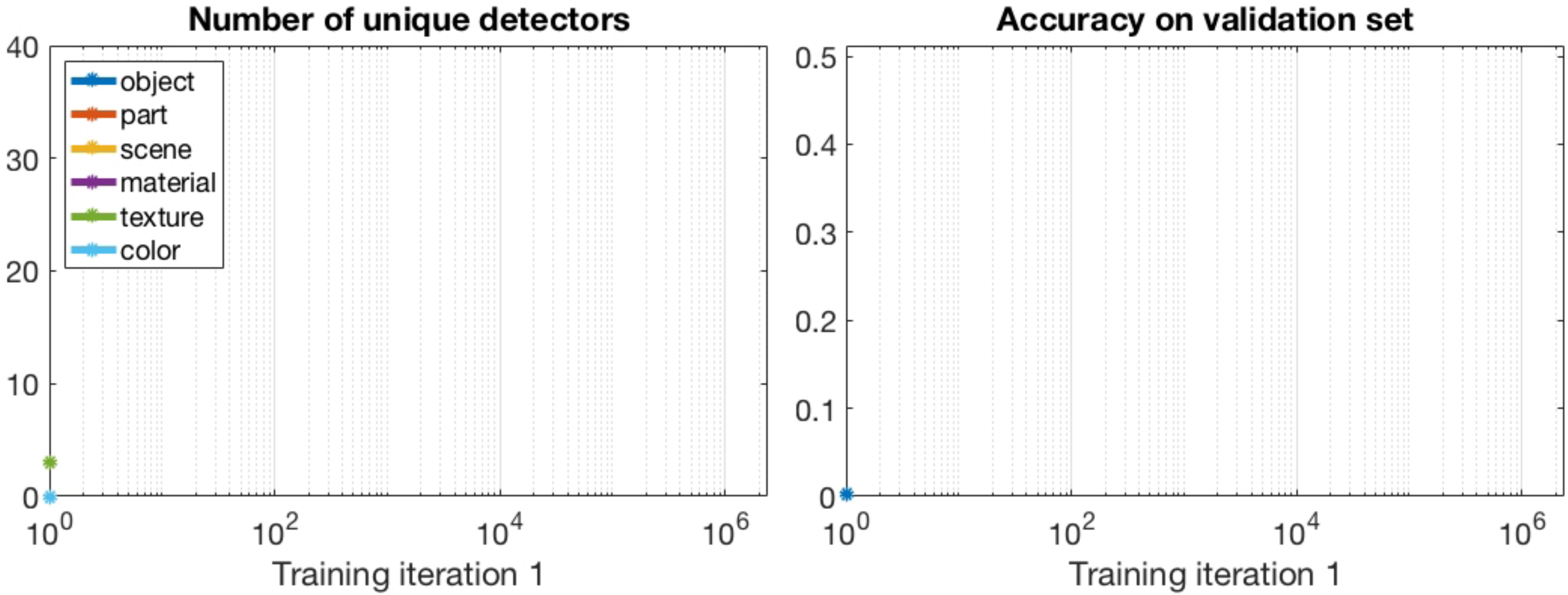


conv5 unit 15: banded (texture) IoU=0.13

conv5 unit 159: tree (object) IoU=0.039

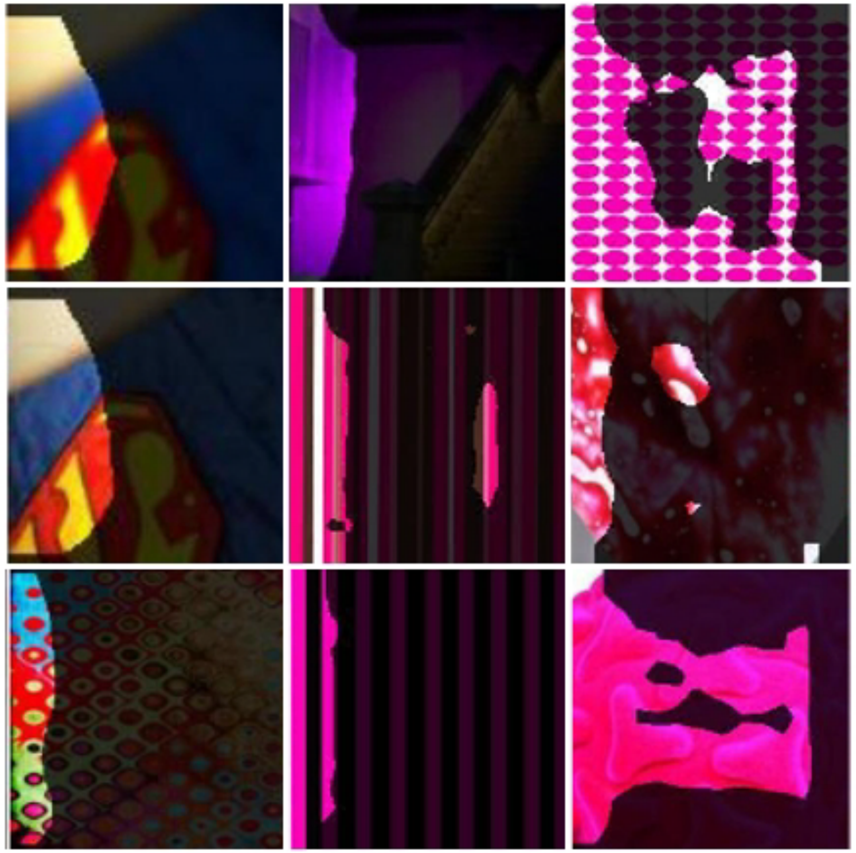
conv5 unit 210: head (part) IoU=0.038

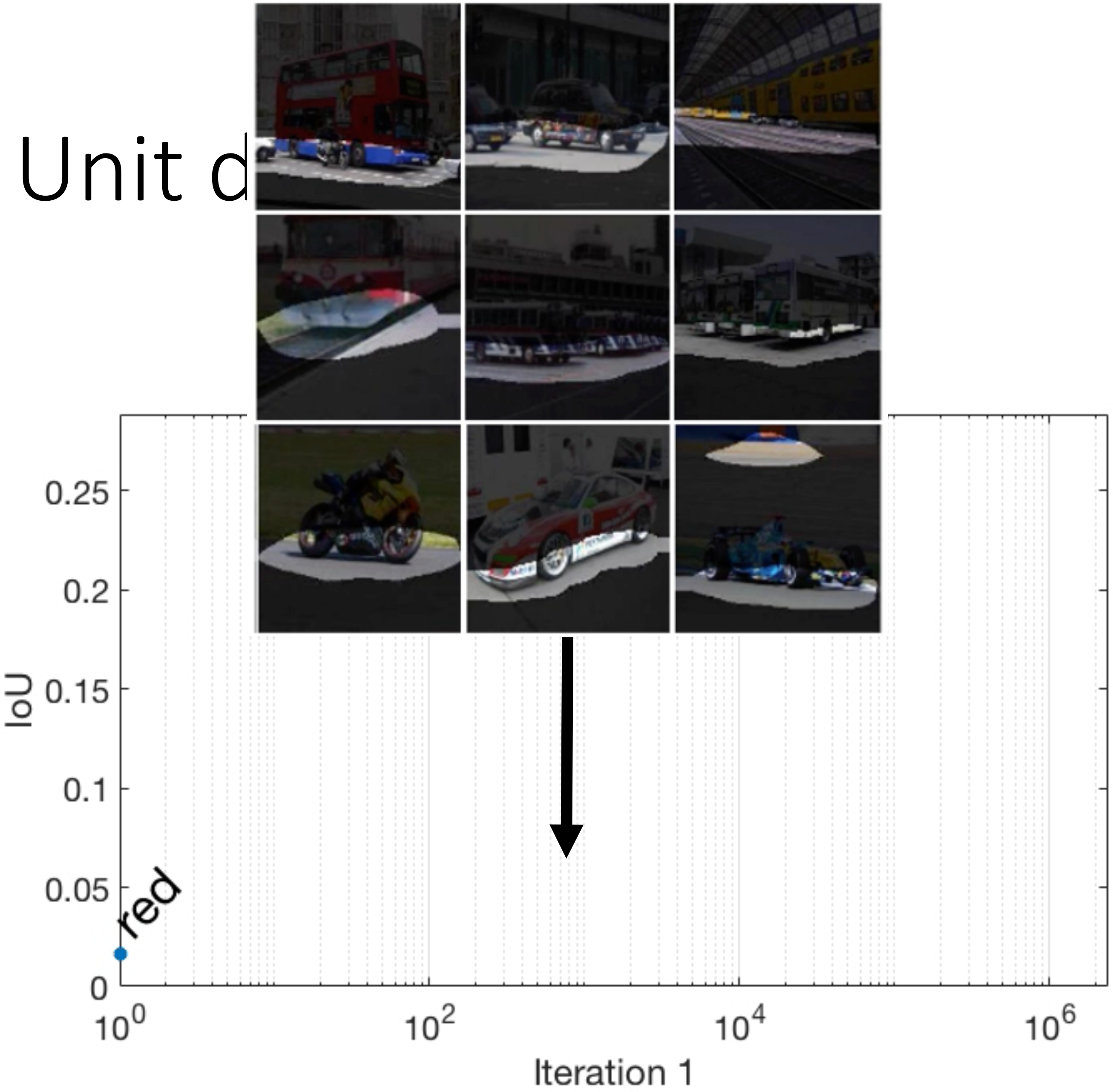
Emergence of Interpretable Units during Training



Individual Unit d

Unit 23 at conv5 layer

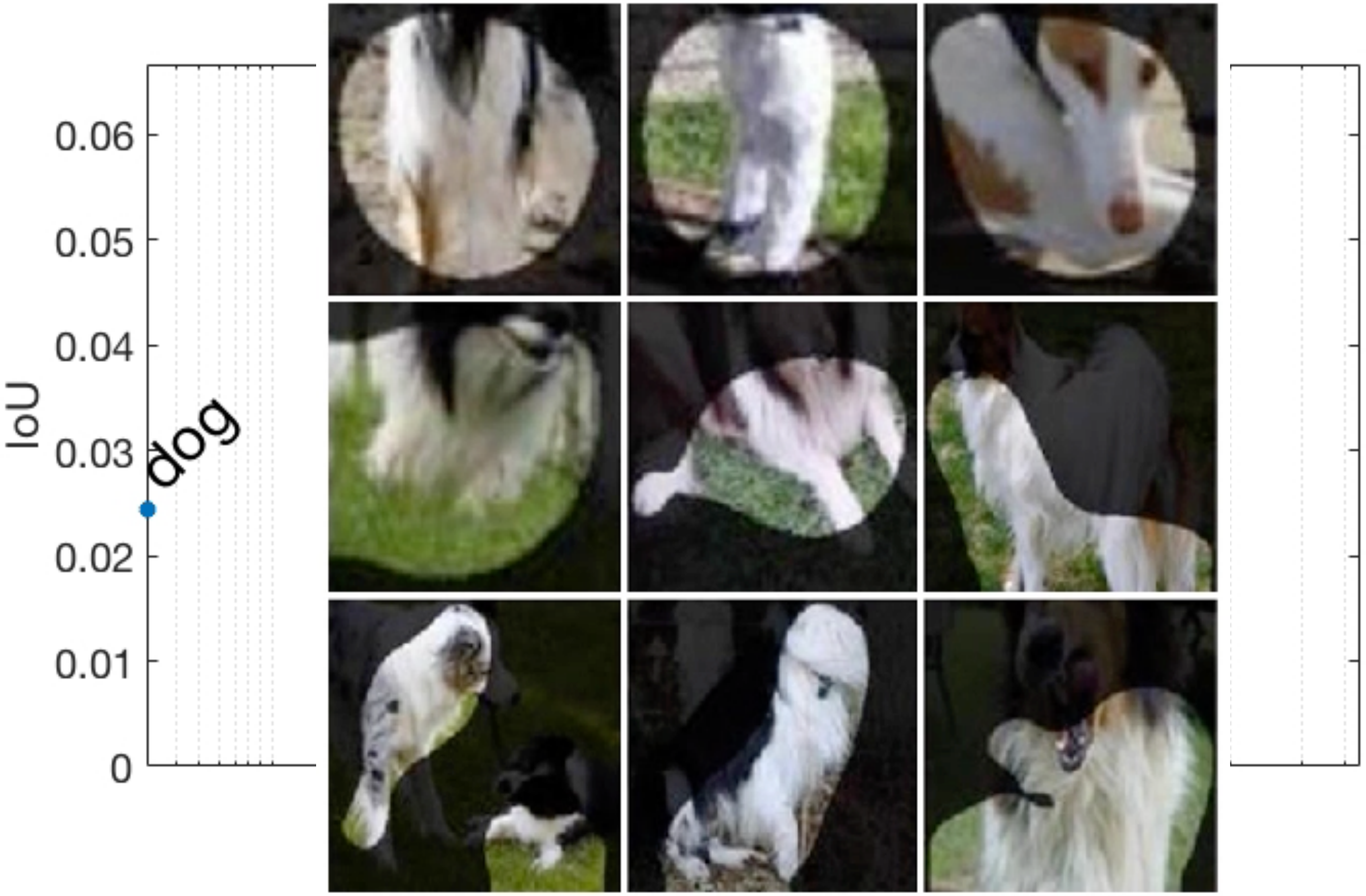




Fine-tuning from ImageNet to Places

Unit 8 at conv5 layer

Before fine-tuning



Fine-tuning from ImageNet to Places

Unit 52 at conv5 layer

Before fine-tuning

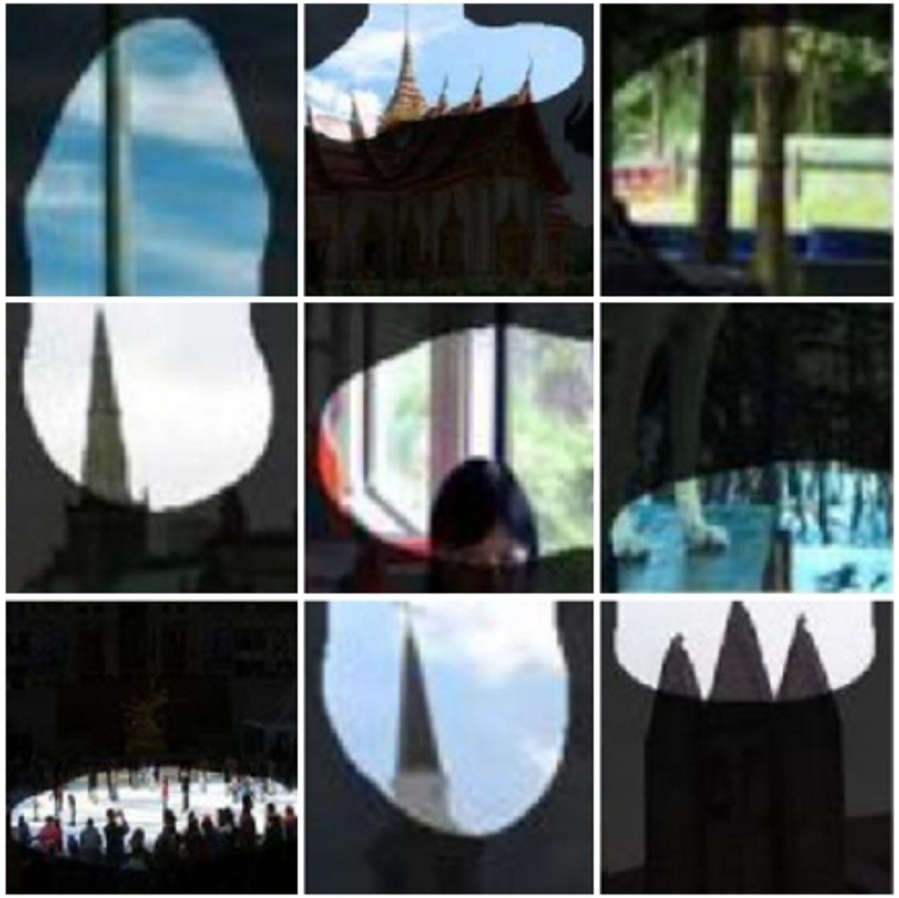
Fine-tuning from Places to ImageNet

Unit 35 at conv5 layer

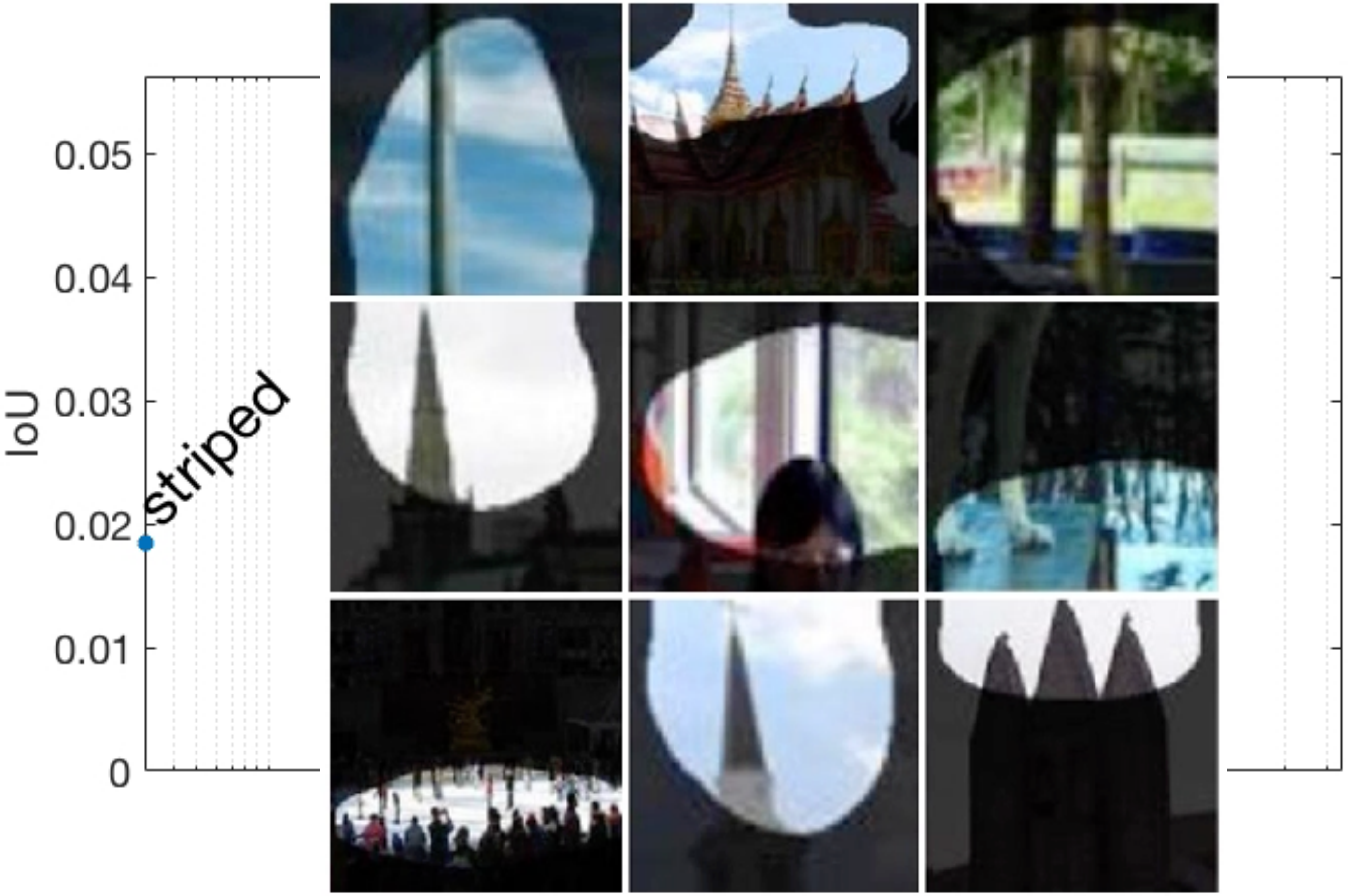
Before fine-tuning

Fine-tuning from Places to ImageNet

Unit 103 at conv5 layer

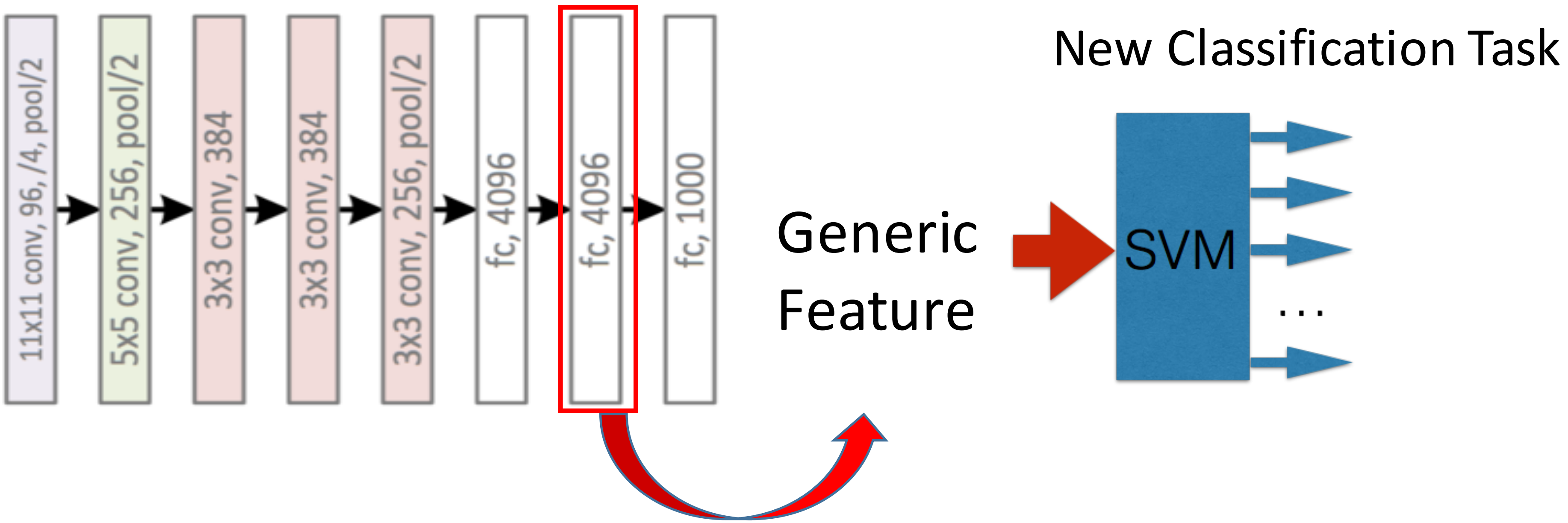


Before fine-tuning

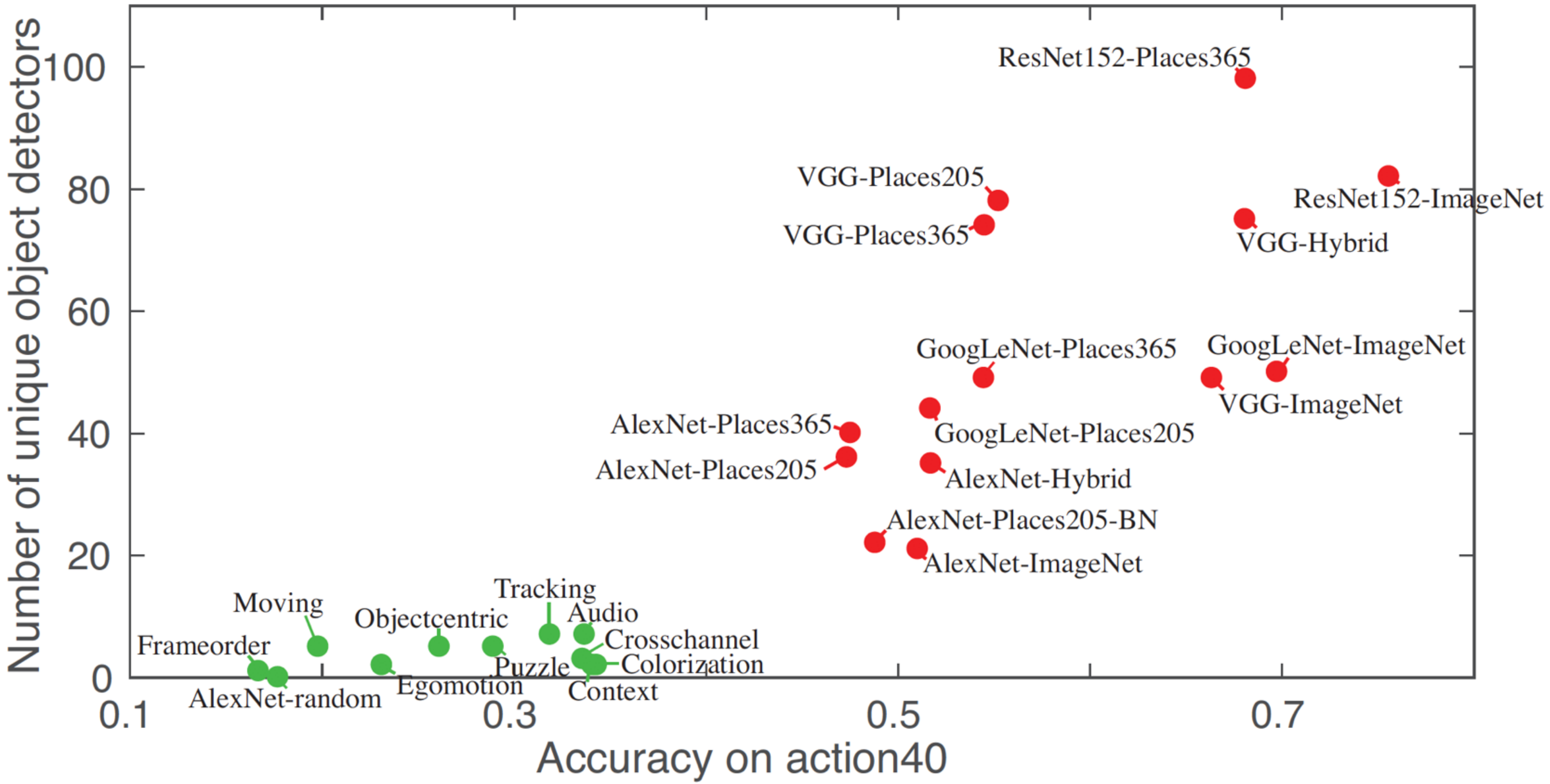


Explainable Deep Features

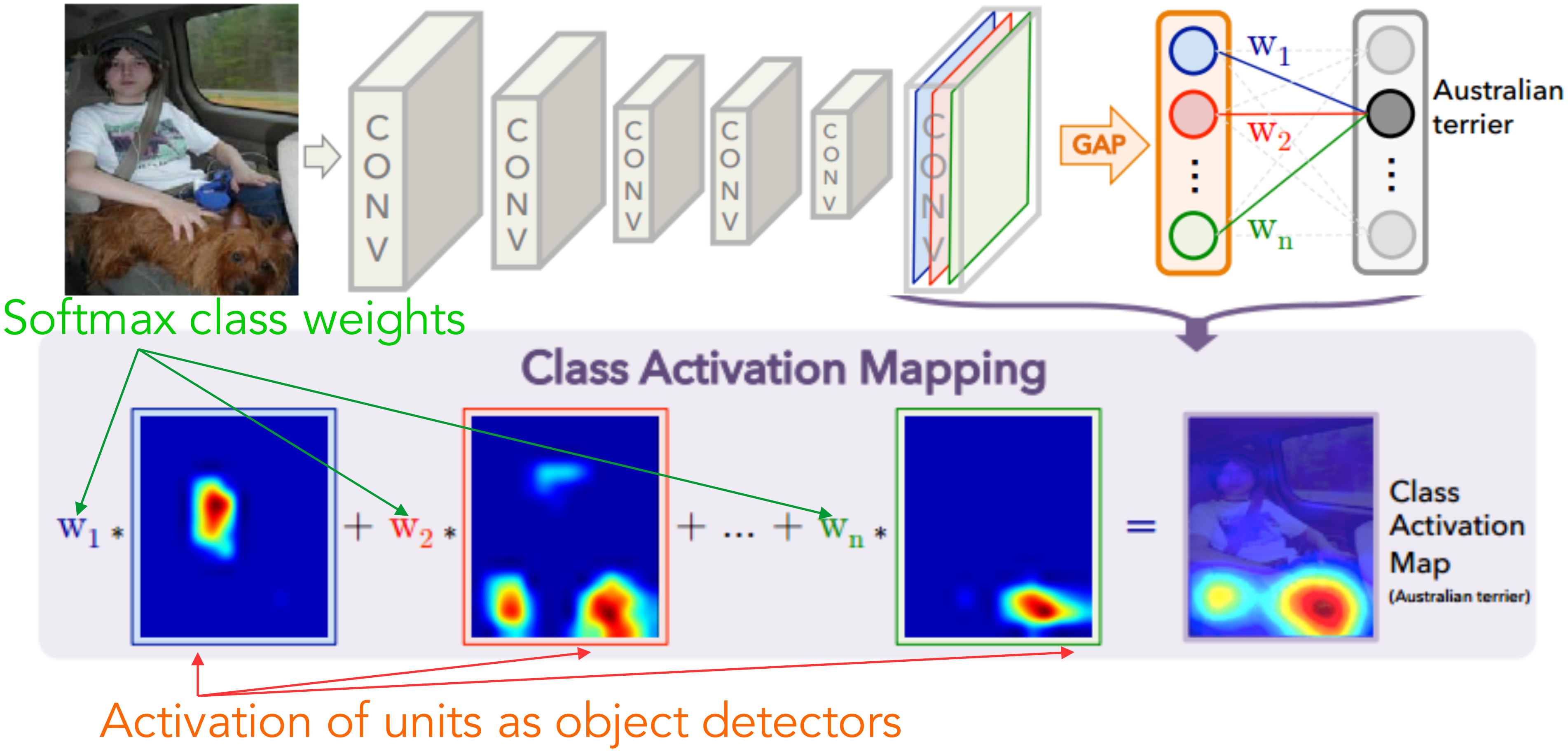
Activations from CNN as generic visual feature



Deep features as generic visual descriptor



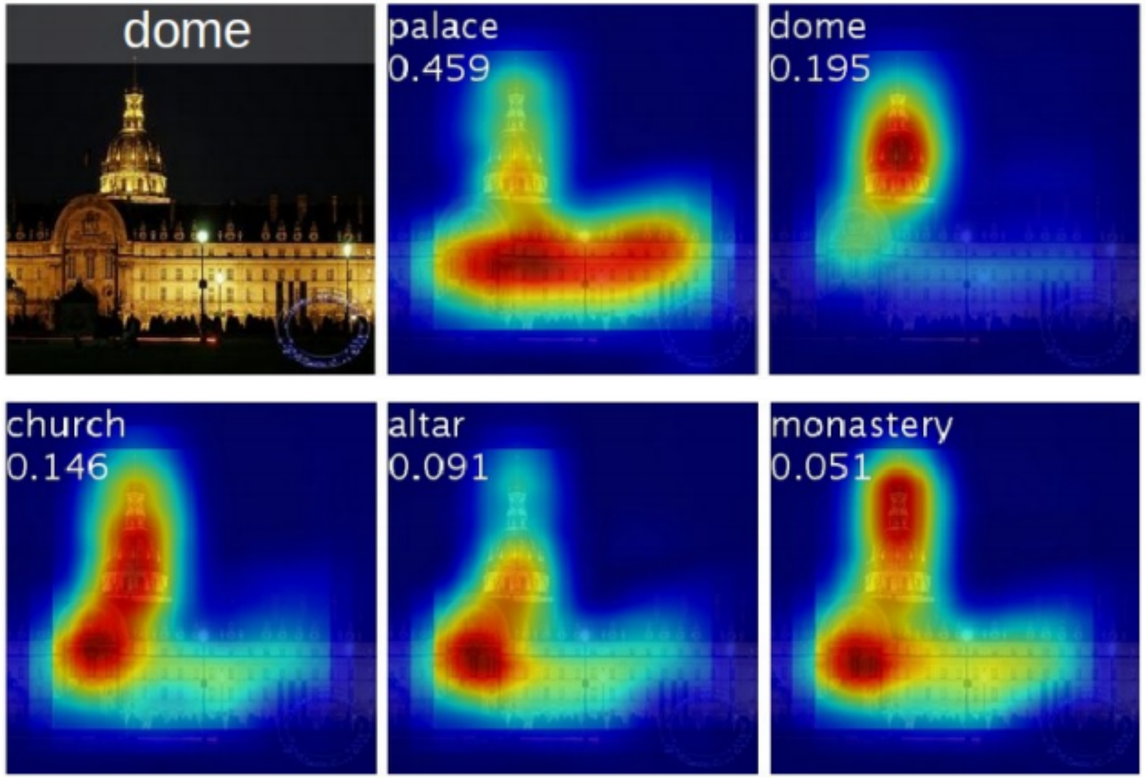
Explaining the Output



Zhou et al. **Learning Deep Features for Discriminative Localization.** CVPR *2016*

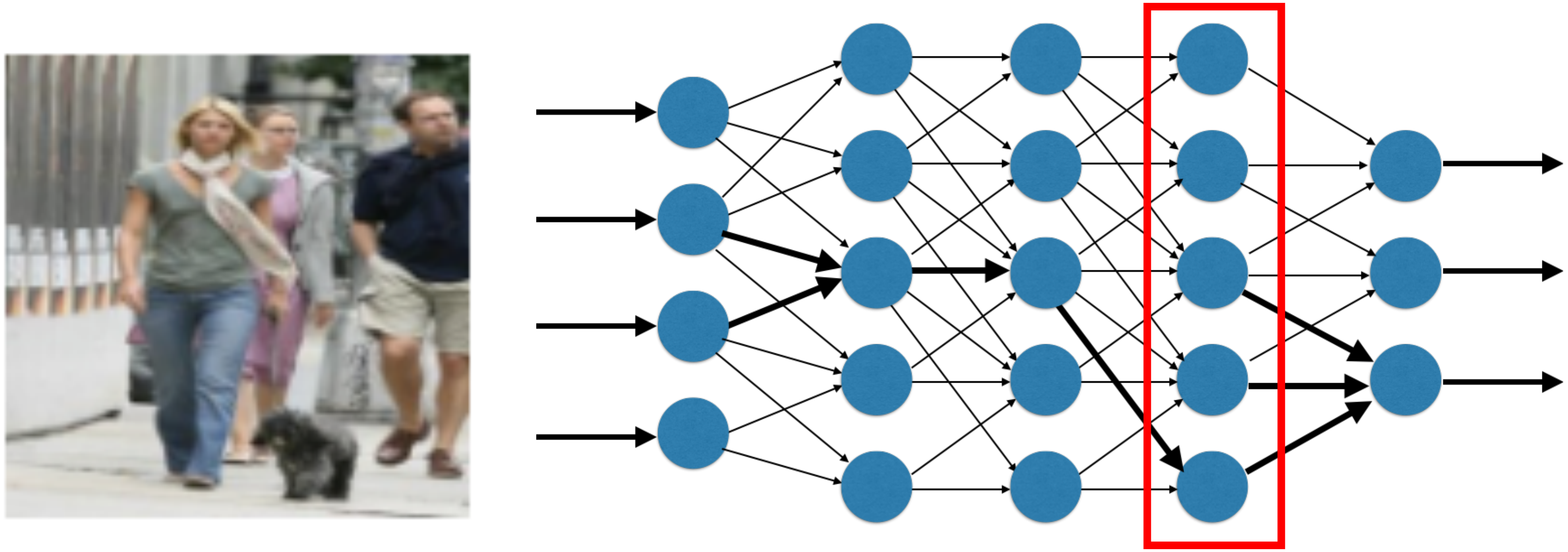
Explaining the Output

- Class Activation Maps (CAM) for the top5 predictions:
  palace, dome, church, altar, monastery



Zhou et al. Learning Deep Features for Discriminative Localization. CVPR *2016*

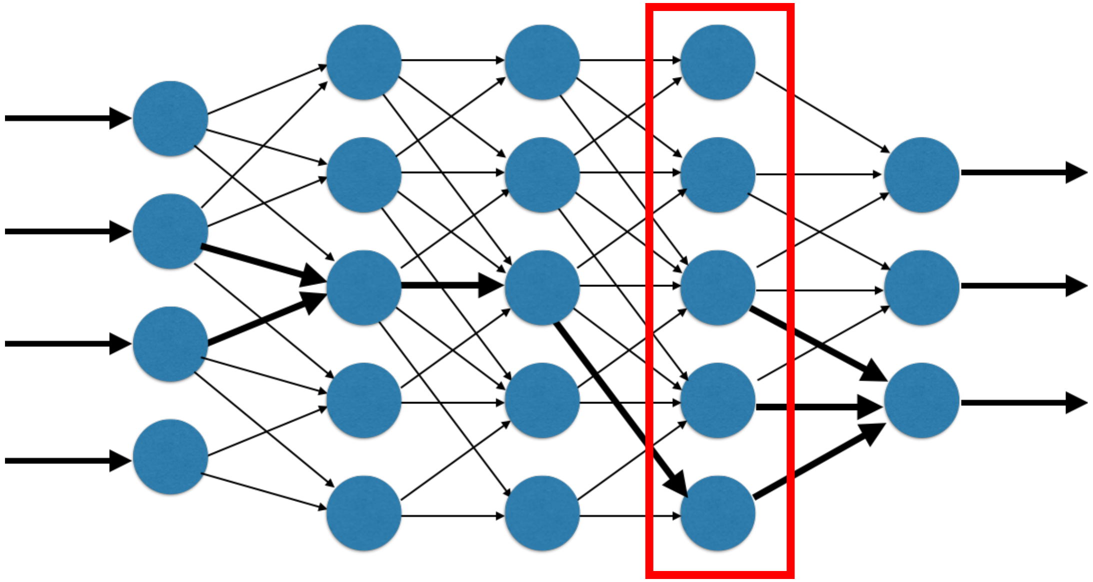
Explaining the Output by Unit Interpretations



Walking the dog

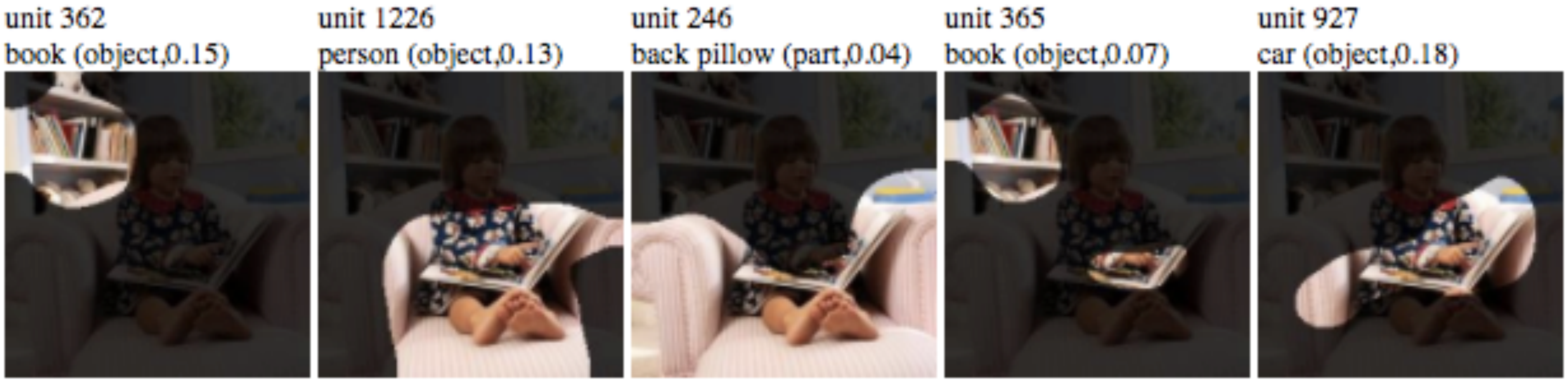
Top activated units

Explaining the Output by Unit Interpretations

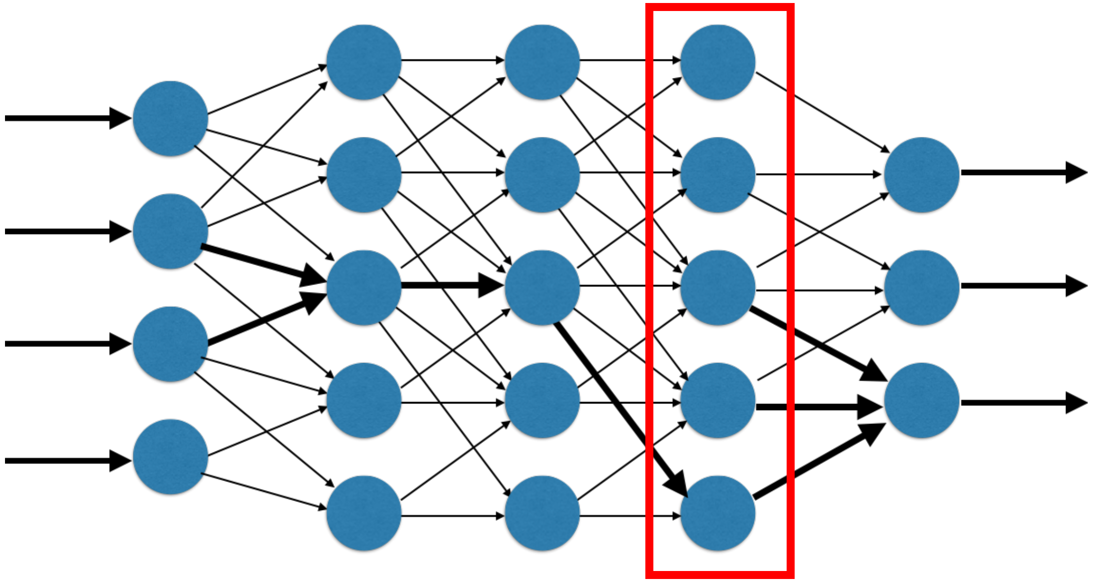


Reading

Top activated units



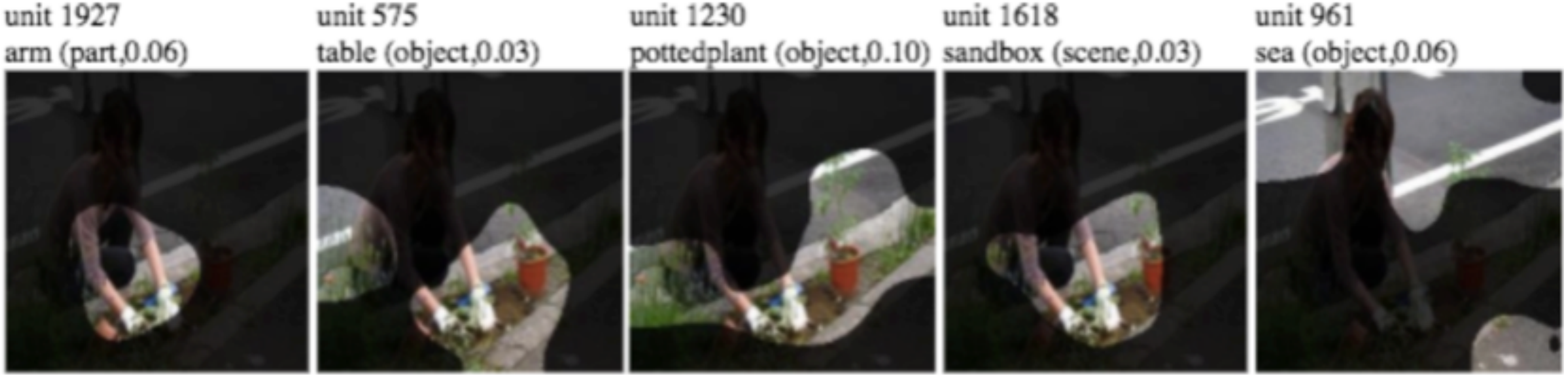
Explaining the Output by Unit Interpretations



Correct label: Gardening

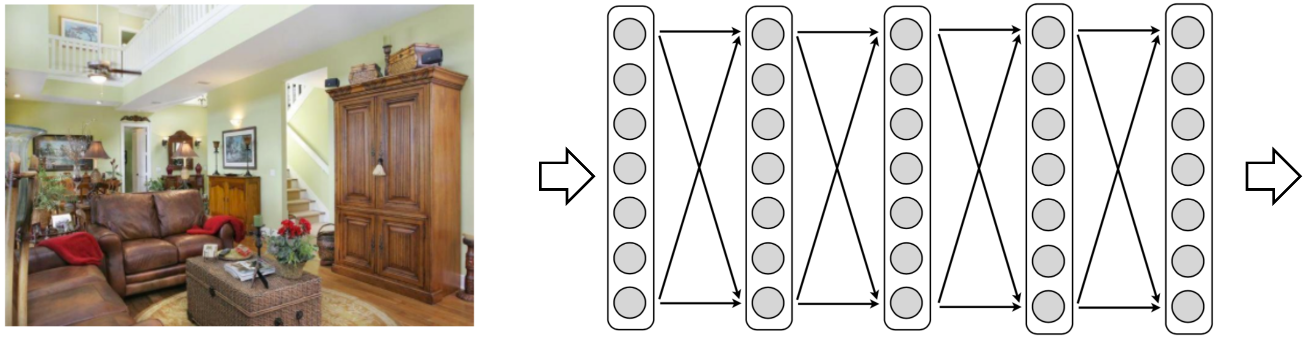
Cutting vegetables

Top activated units



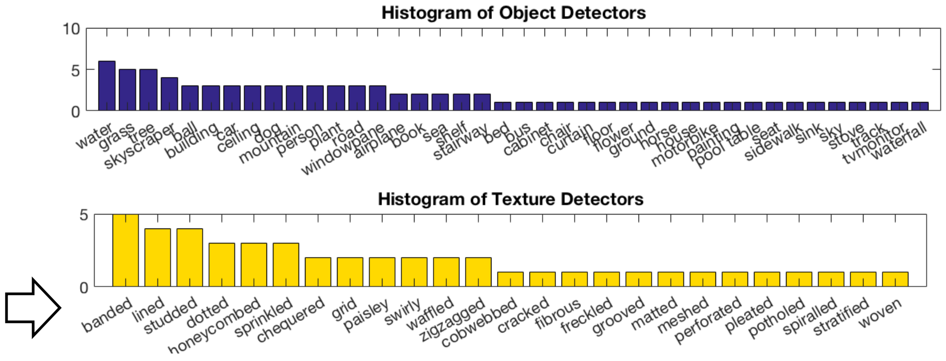
Code and more visualizations are at http://netdissect.csail.mit.edu

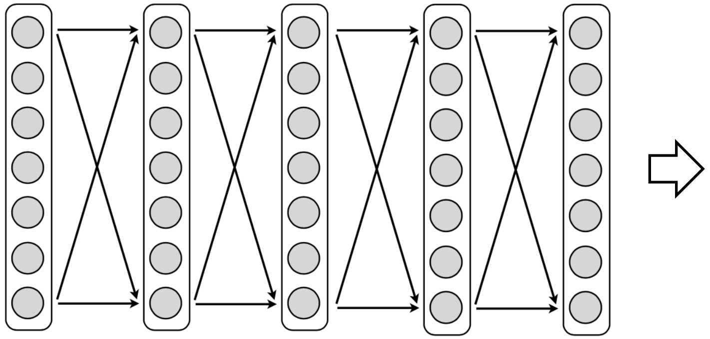
Conclusion



Living room
Kitchen
Coast
Theater
…

Interpretability Report





Network Dissection

unit 79 car, IoU=0.13

unit 107 road, IoU=0.15