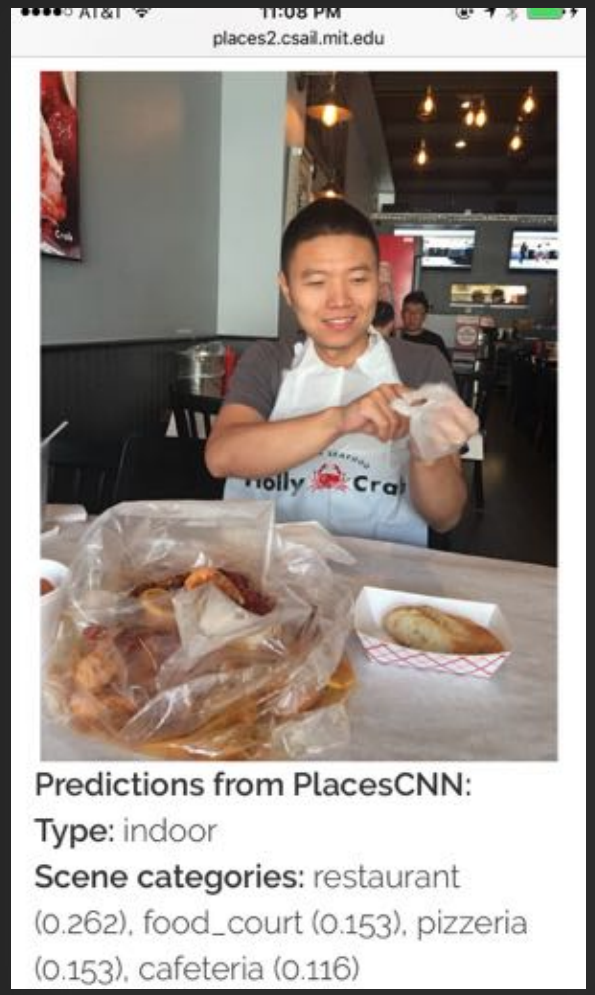
Supervised and Self-supervised Representation Learning

Bolei Zhou

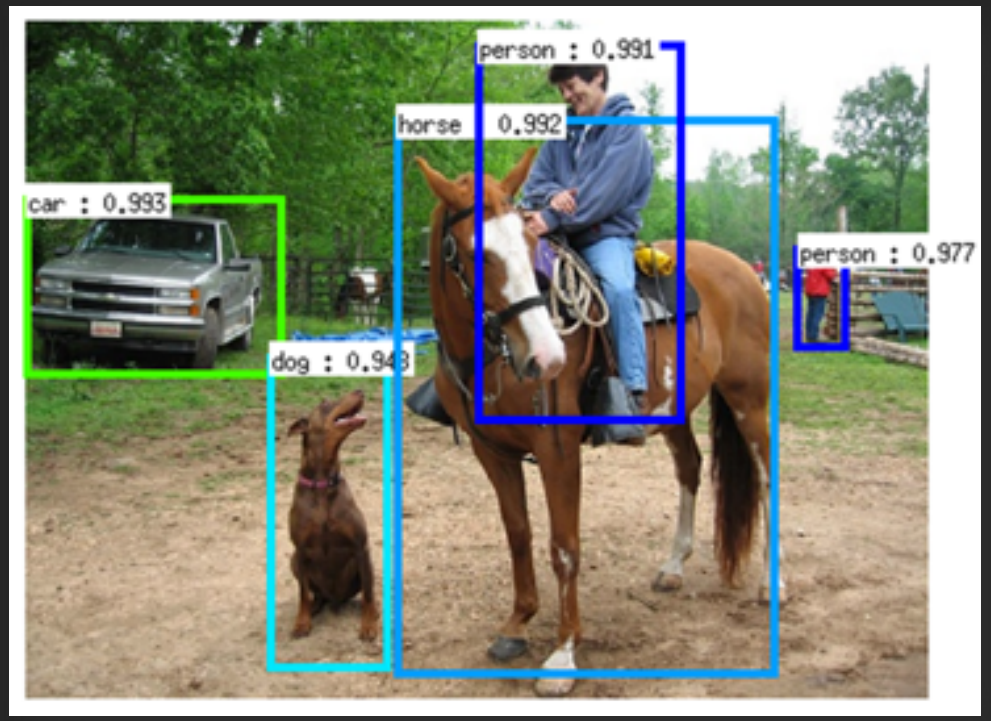
The Chinese University of Hong Kong

Deep Neural Networks for Computer Vision

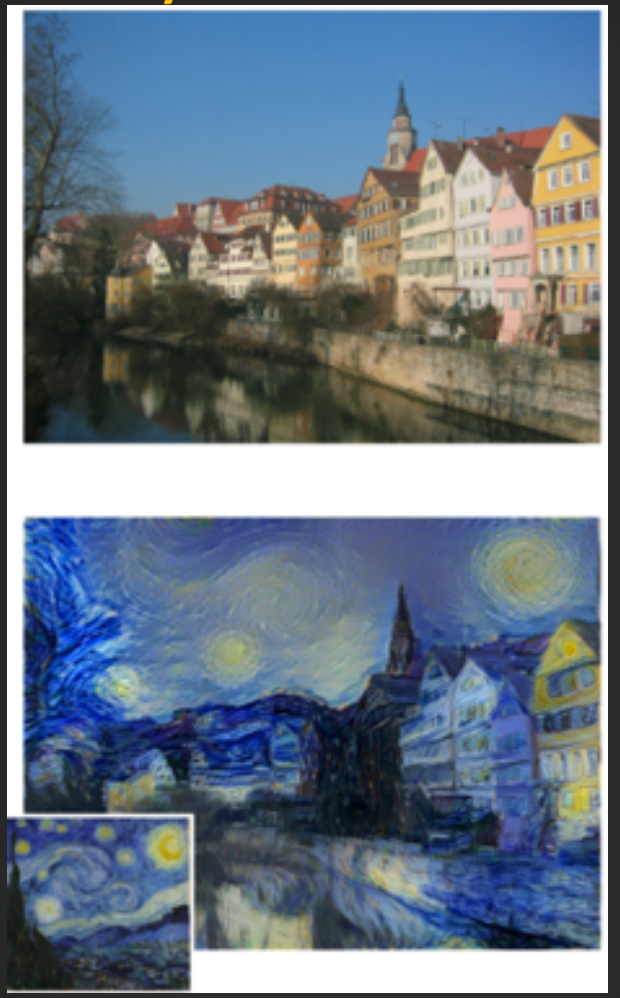
Image recognition



Object detection

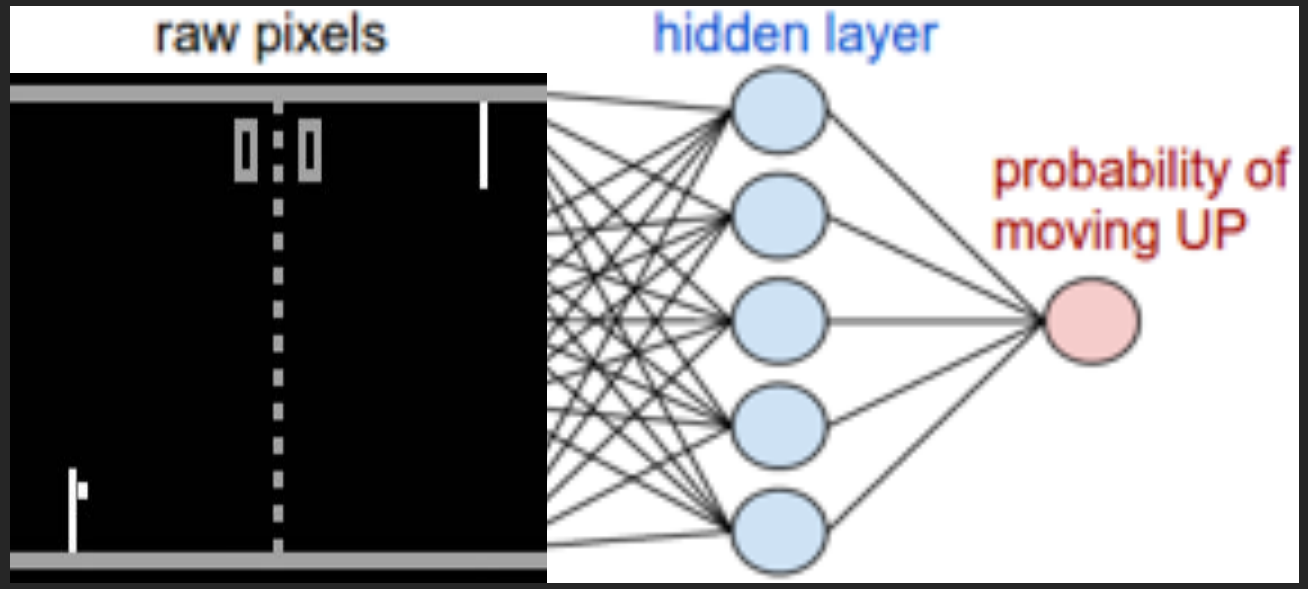


Style transfer

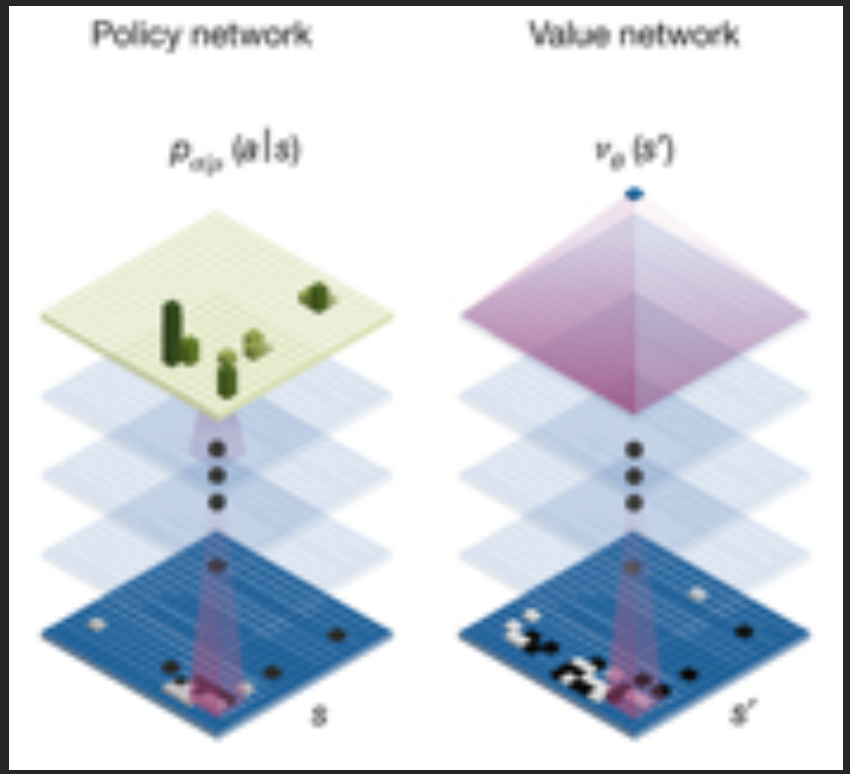


Deep Neural Networks for RL Agents

Playing Atari game

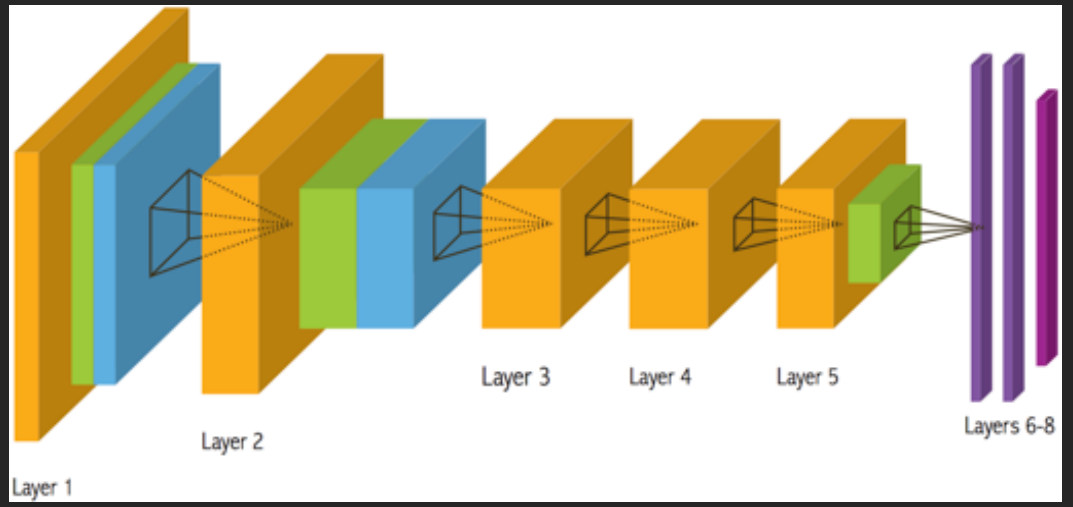


AlphaGo



Representations from Supervised Learning

CNN trained Places for Scene Recognition



365 classes

Bedroom

...

Training data: 2.6 million images from Places Database (Zhou et al.)

places

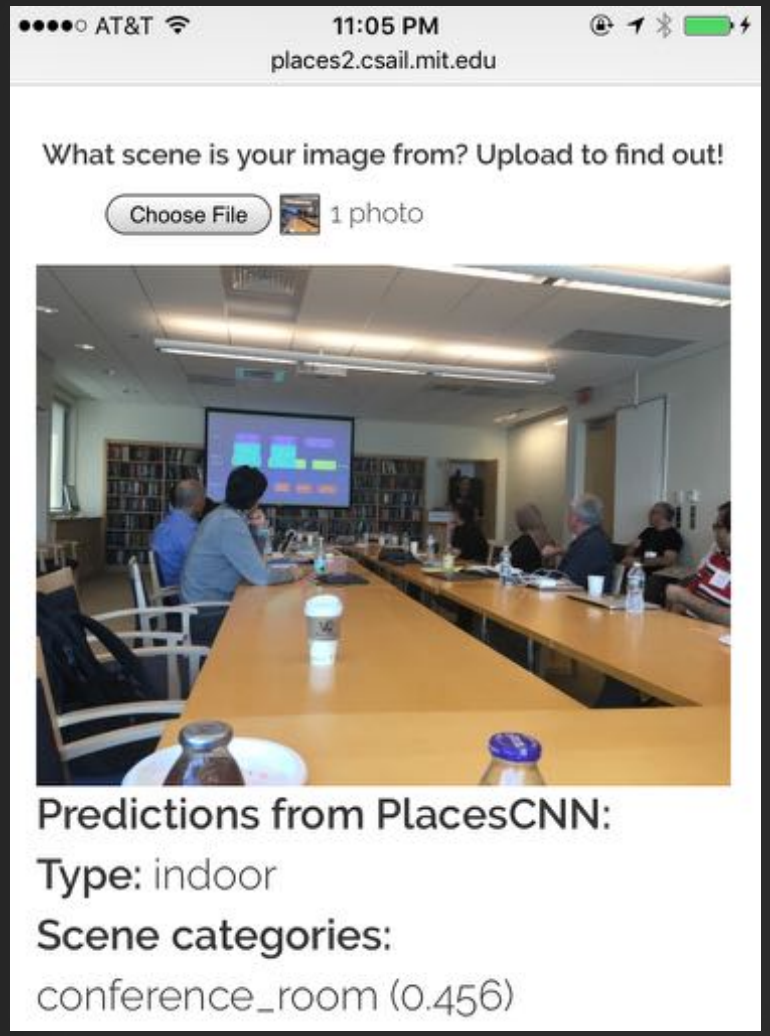
spare bedroom

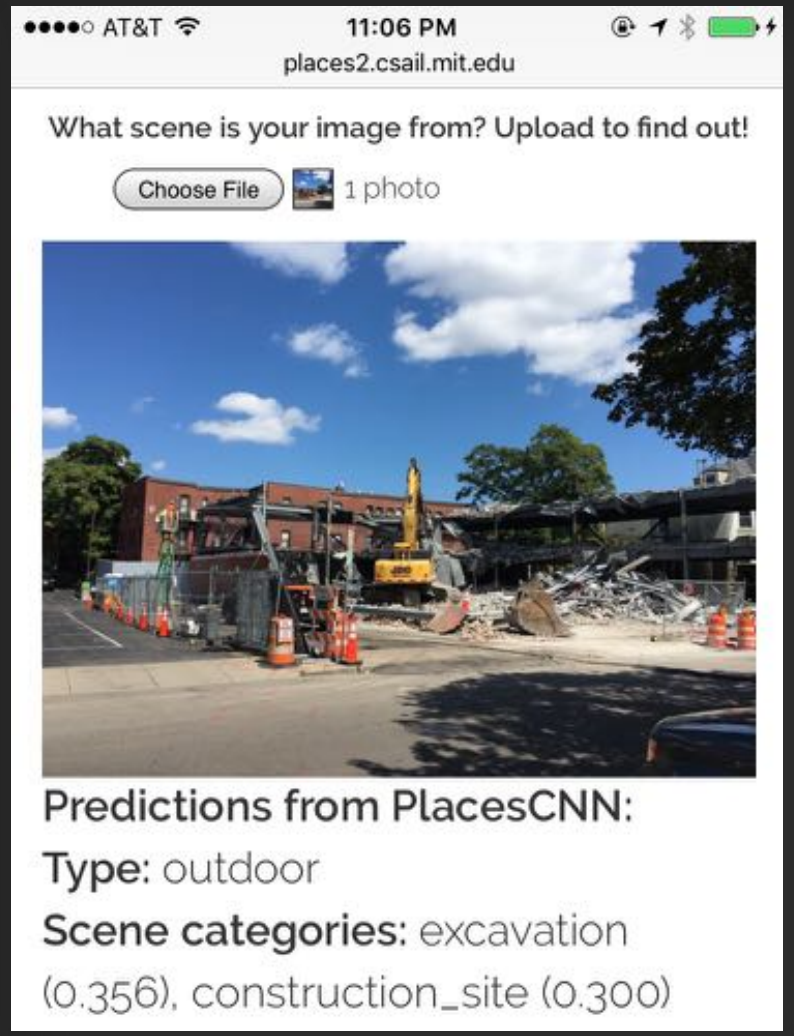
messy bedroom

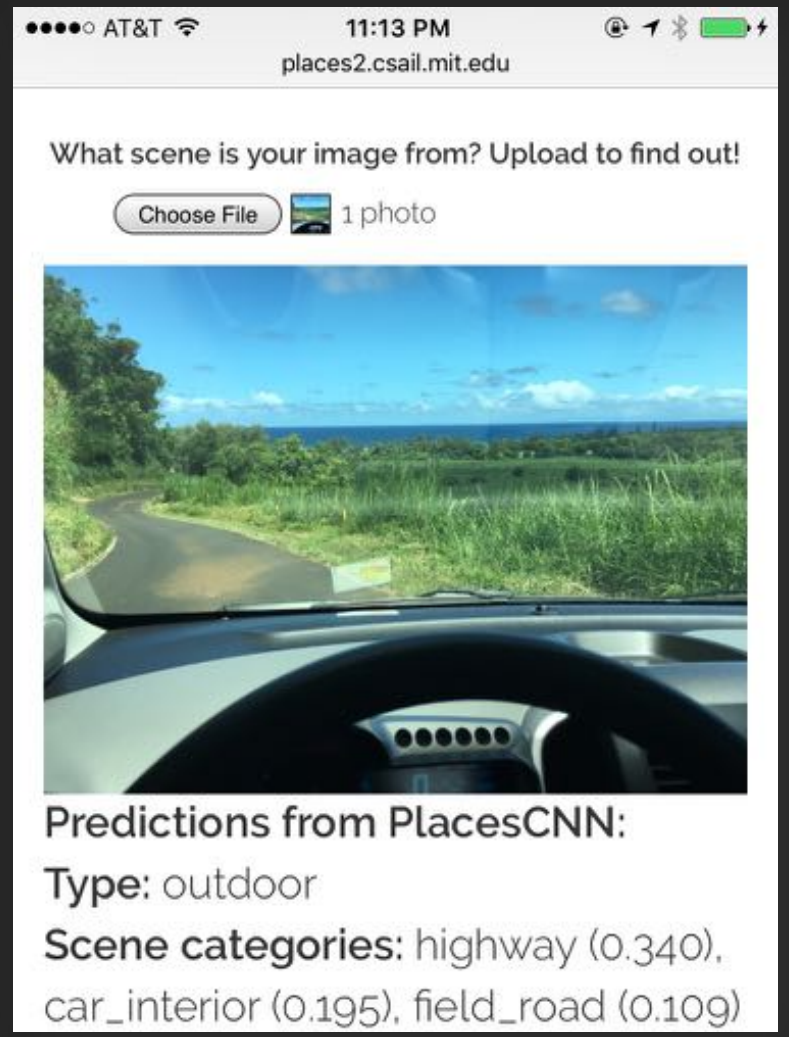
teenage bedroom

romantic bedroom

PlacesCNN works so well







http://places2.csail.mit.edu/demo.html

PlacesCNN works so well

places2.csail.mit.edu

Predictions from PlacesCNN:

Type: outdoor

Scene categories: field/cultivated (0.839)

places2.csail.mit.edu

Predictions from PlacesCNN:

Type: outdoor

Scene categories: lake/natural (0.387), river (0.102)

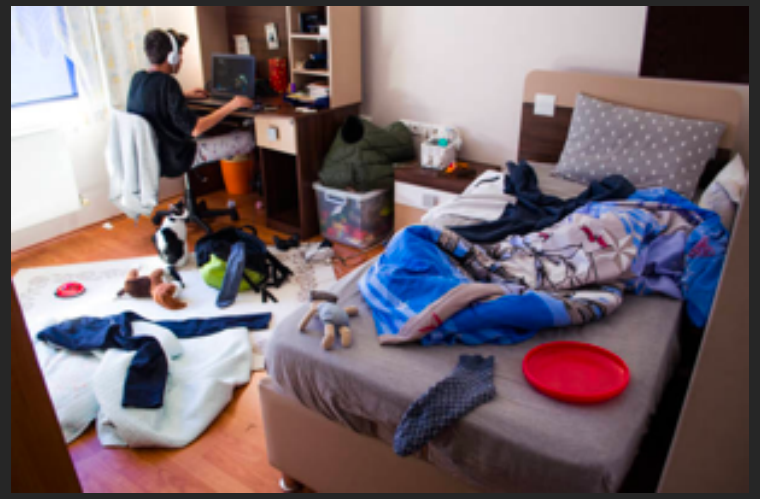
places2.csail.mit.edu

Predictions from PlacesCNN:

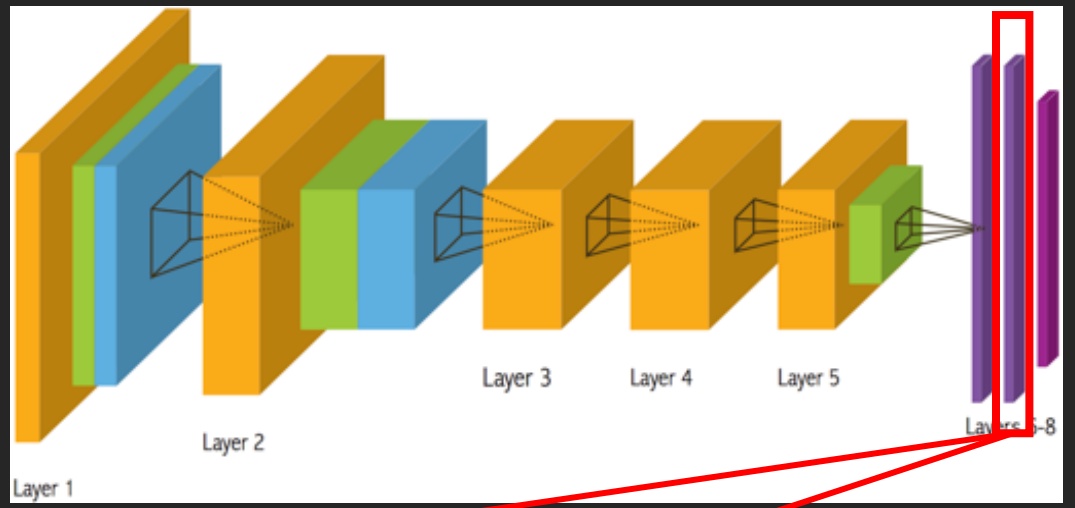
Type: indoor

Scene categories: shower (0.135), bathroom (0.133)

Deep Representations as Generic Visual Features



CNN trained Places for Scene Recognition

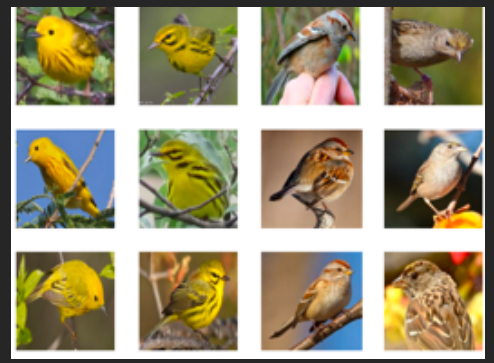


Bedroom

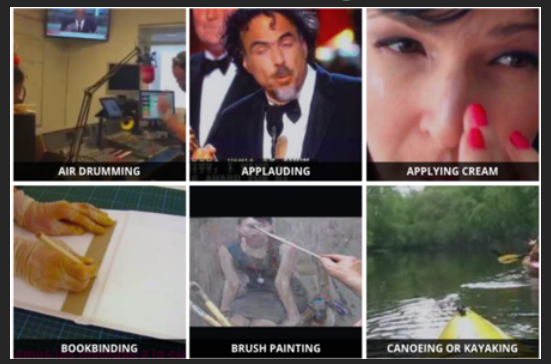
…

1024-dimensional vector

Bird classification



Action recognition



Deep Feature + Linear SVM work well for transferred tasks!

Why deep features are transferrable?

What have been learned inside?

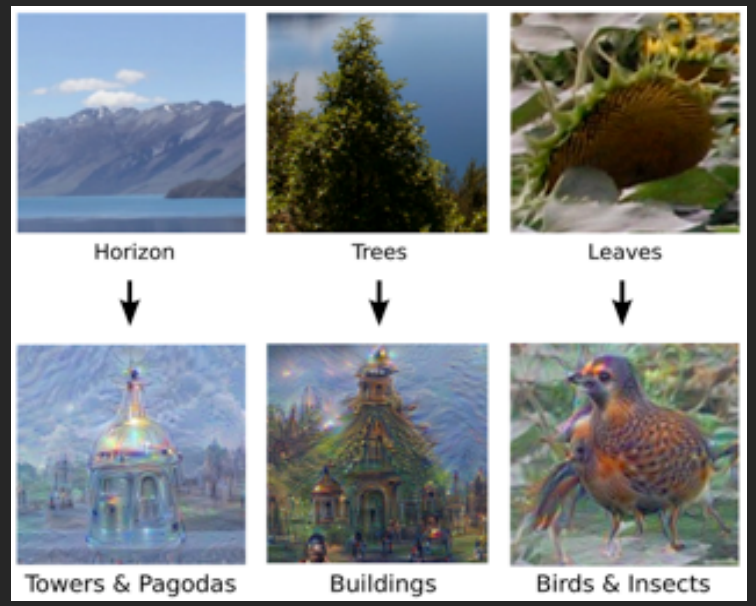
Work on Visualizing Networks

Deconvolution

Zeiler et al., ECCV 2014.

Back-propagation

Simonyan et al., ICLR 2015

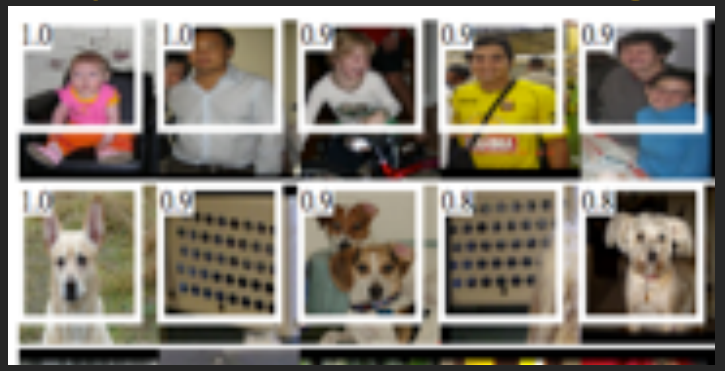


Inceptionism. Google Blog. June 2015

Feature inversion

Mahendran et al, CVPR 2015

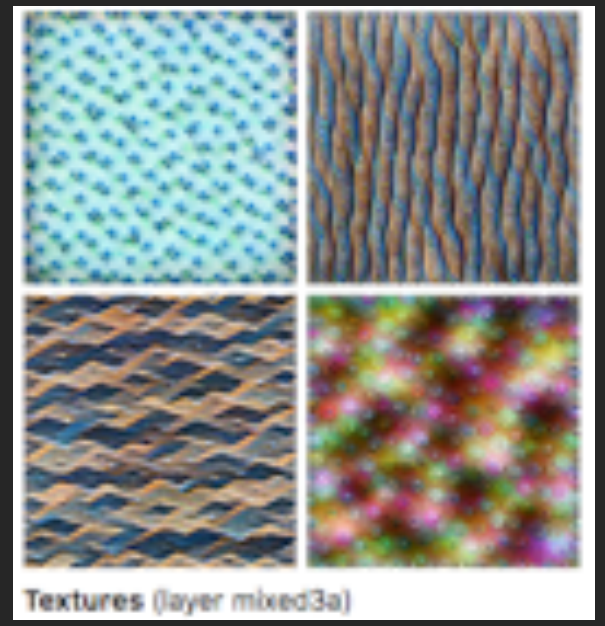
Top activated images

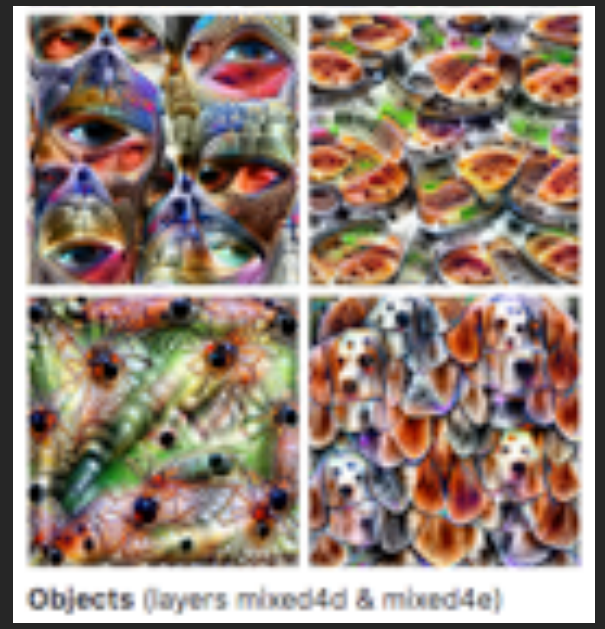


Girshick et al., CVPR 2014

Gradient-based Visualization

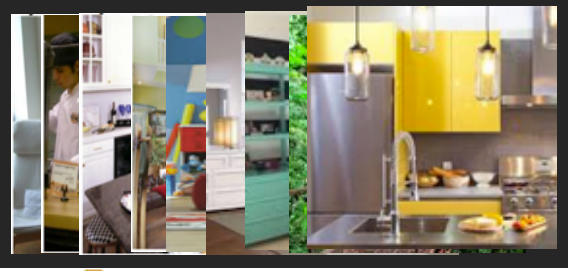
Iteratively use gradient to optimize an image to activate a particular unit

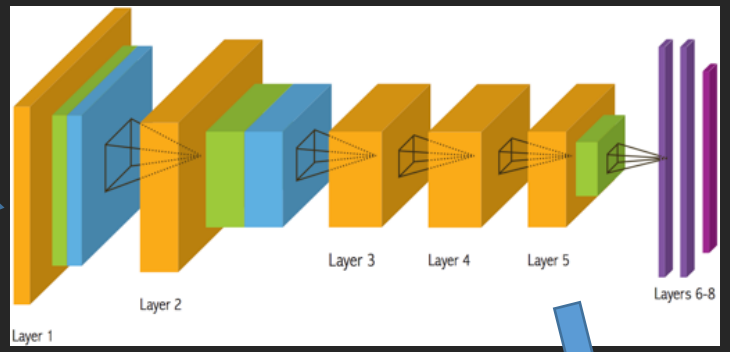


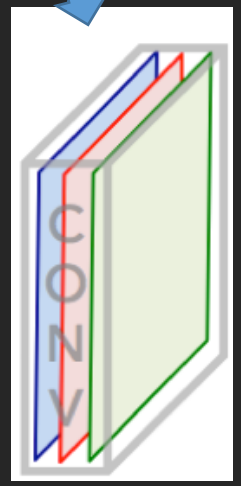


Chris Olah, et al. https://distill.pub/2017/feature-visualization/

Visualizing Internal Units

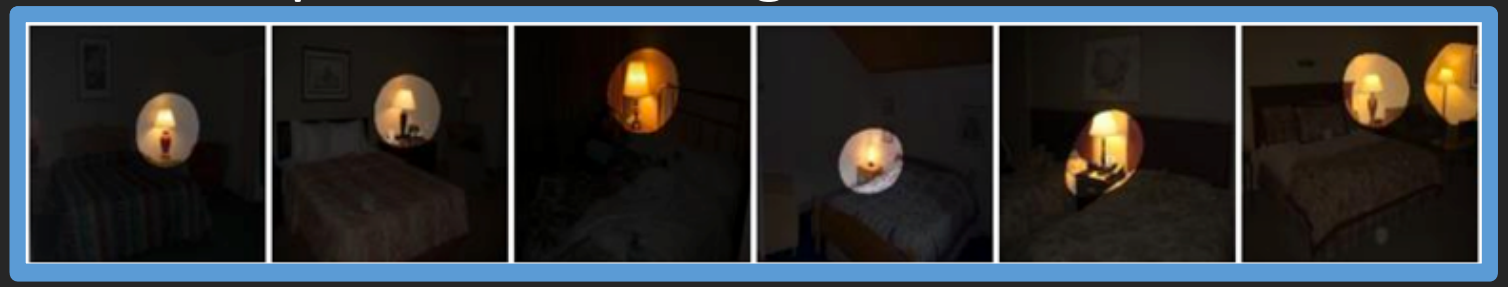




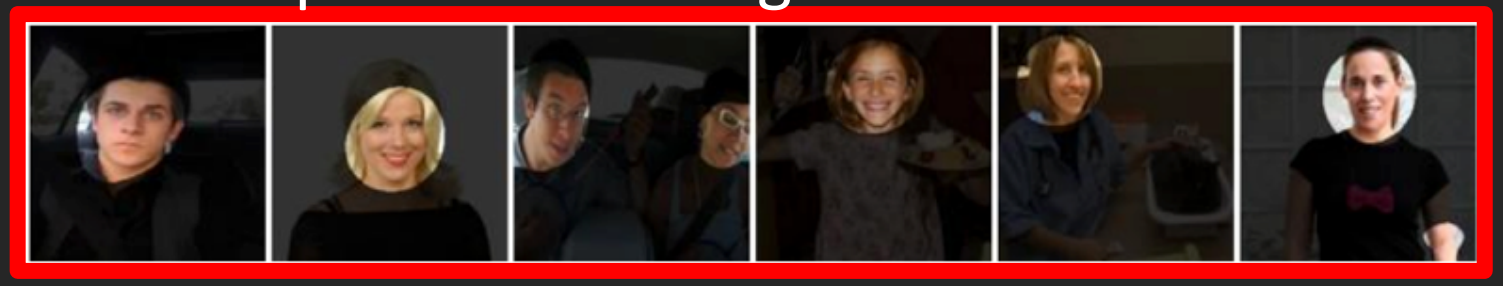


Layer 5

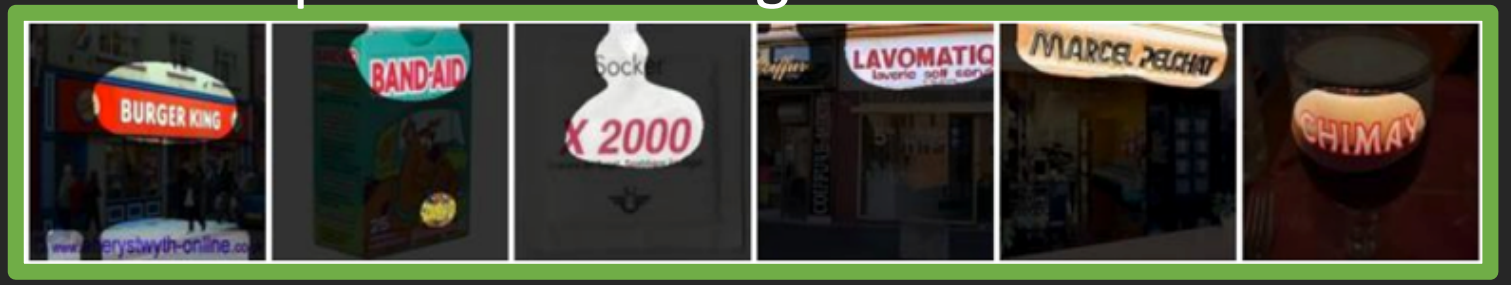
Unit1: Top activated images



Unit2: Top activated images



Unit3: Top activated images



Annotating the Interpretation of Units

Amazon Mechanical Turk

Word/Description to summarize the images:

Lamp

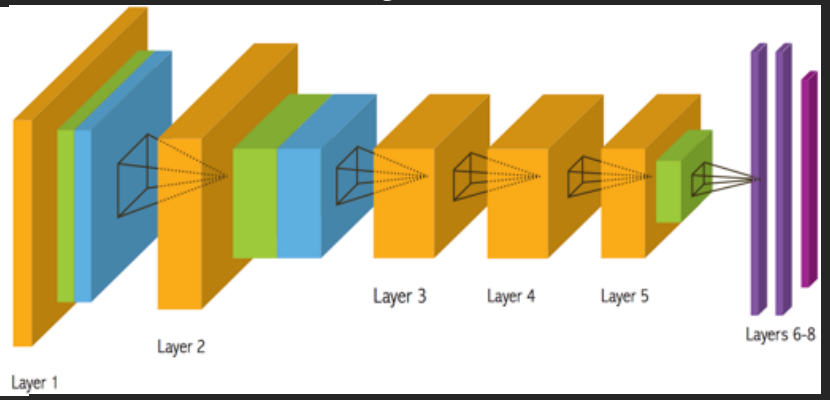
Which category the description belongs to:

- Scene
- Region or surface
- Object
- Object part
- Texture or material
- Simple elements or colors

[Zhou, Khosla, Lapedriza, Oliva, Torralba. ICLR 2015]

Two Recognition Tasks and Two Networks

CNN for Object Classification

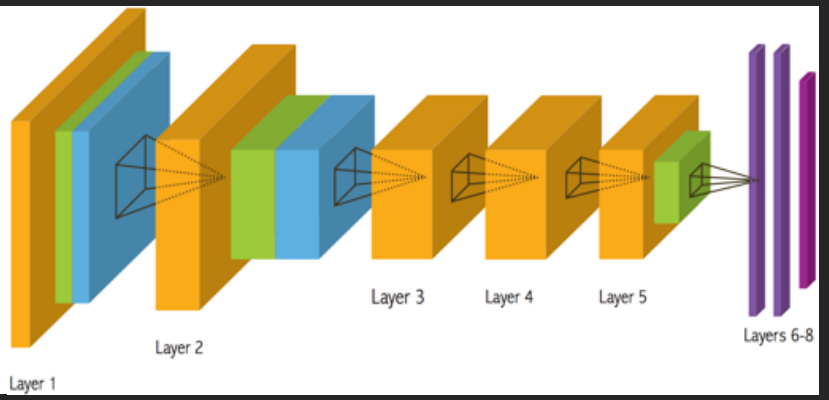


1000 classes

Race car

...

CNN for Scene Recognition



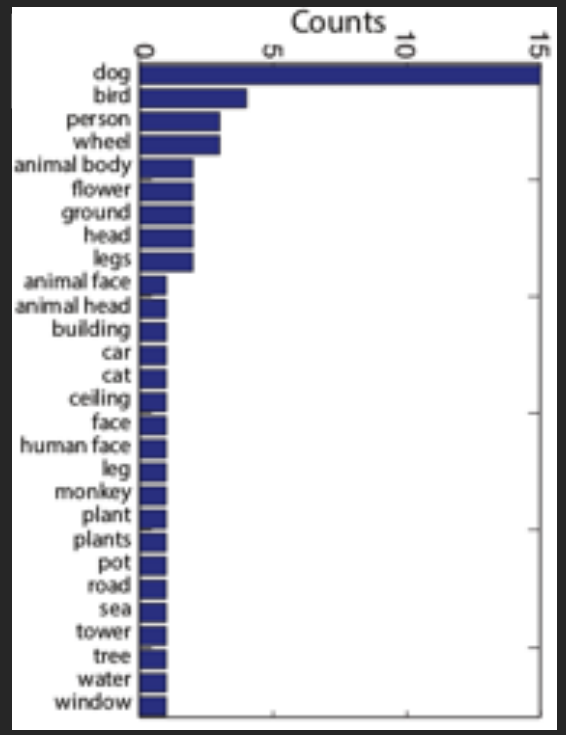
365 classes

Living room

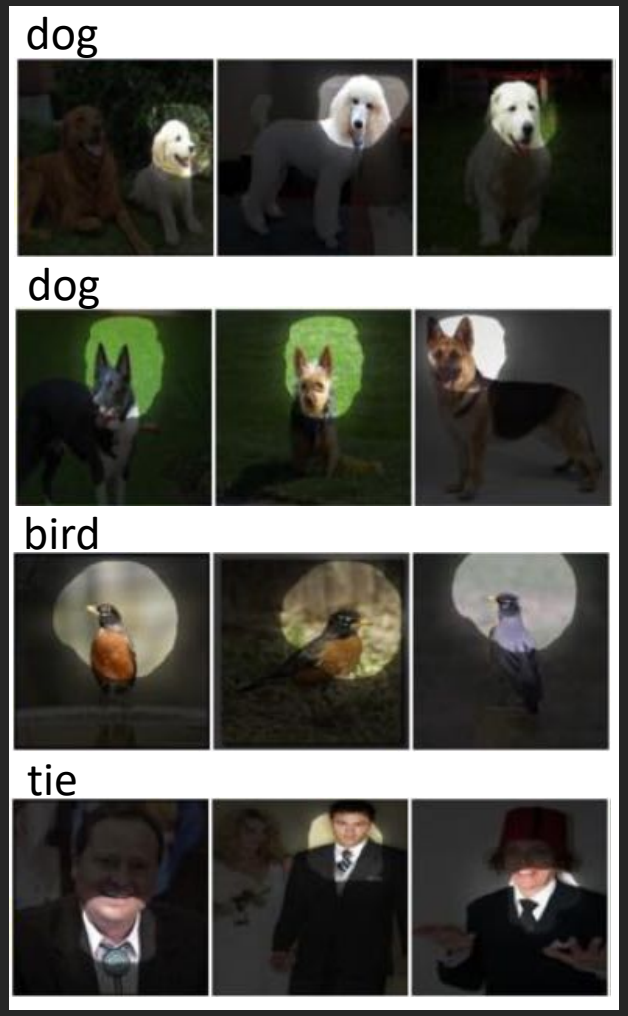
...

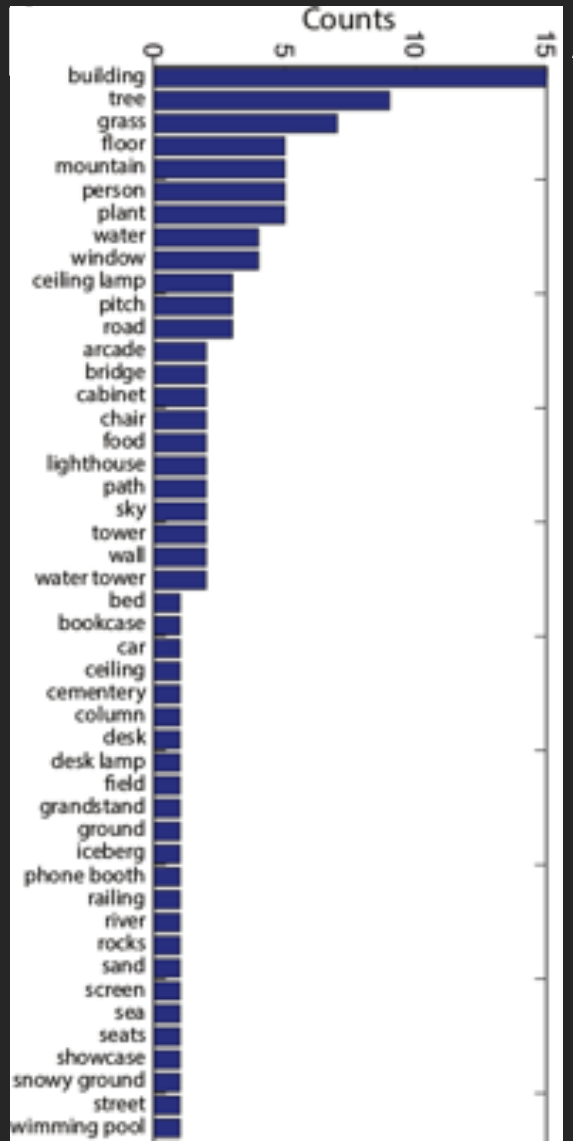
[**Zhou**, Khosla, Lapedriza, Oliva, Torralba. ICLR 2015]

Interpretable Representations for Objects and Scenes

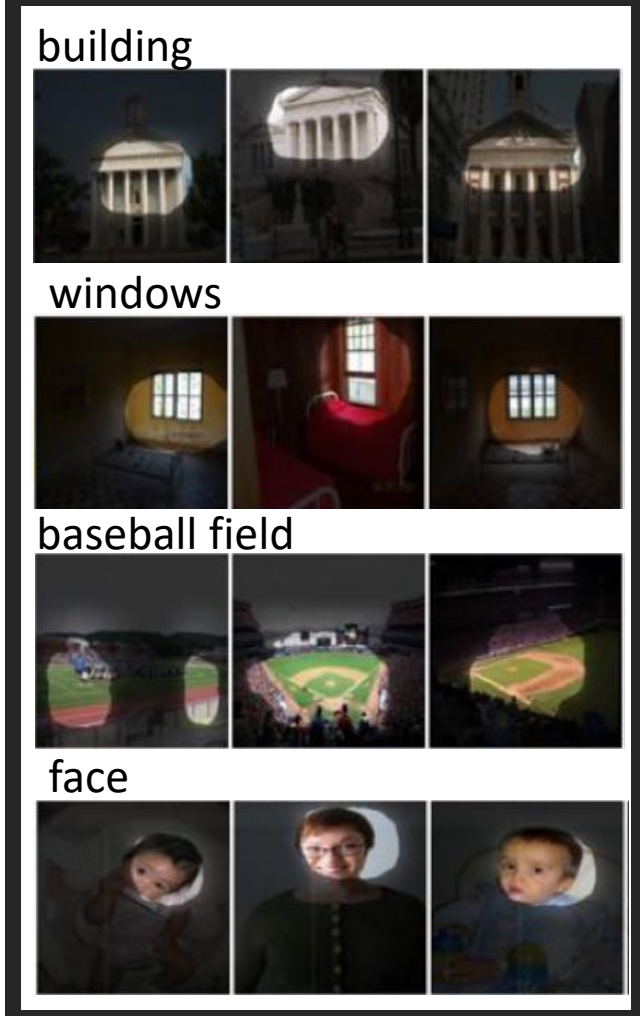


59 units as objects at conv5 of AlexNet on ImageNet





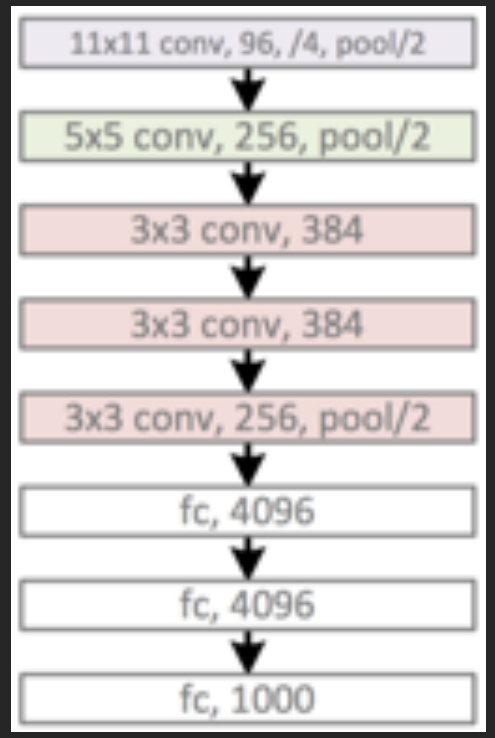
151 units as objects at conv5 of AlexNet on Places



Manually annotating them?

AlexNet: 5 conv layers

~1,000 units



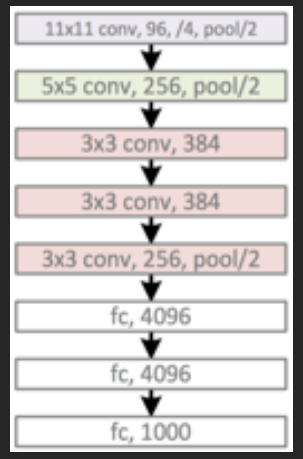
ResNet:

152 conv layers

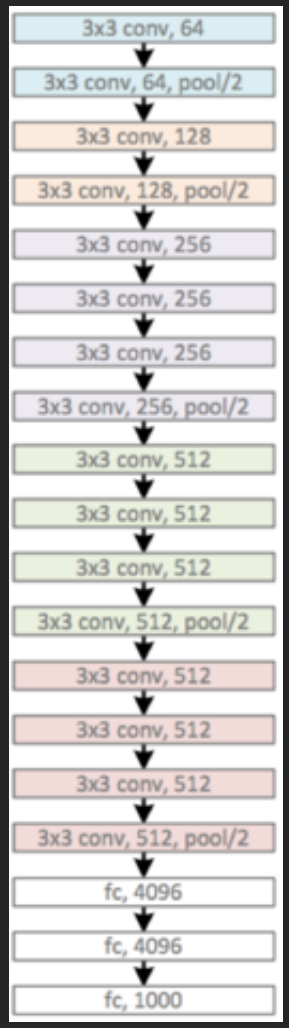
~ 100,000 units

Compare Different Representations of Architectures

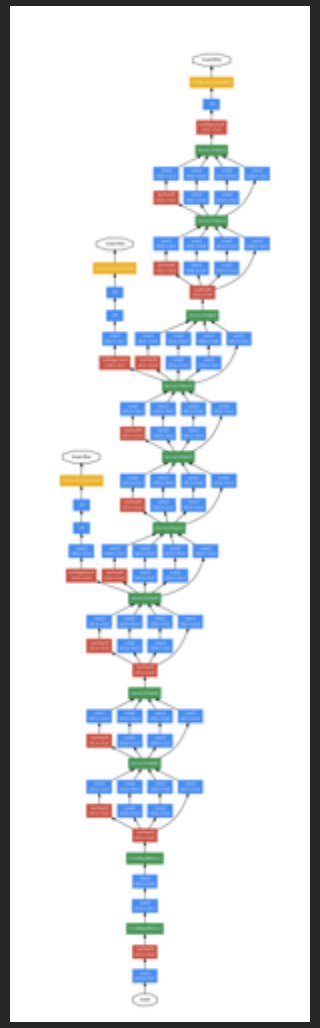
AlexNet



VGG



GoogLeNet



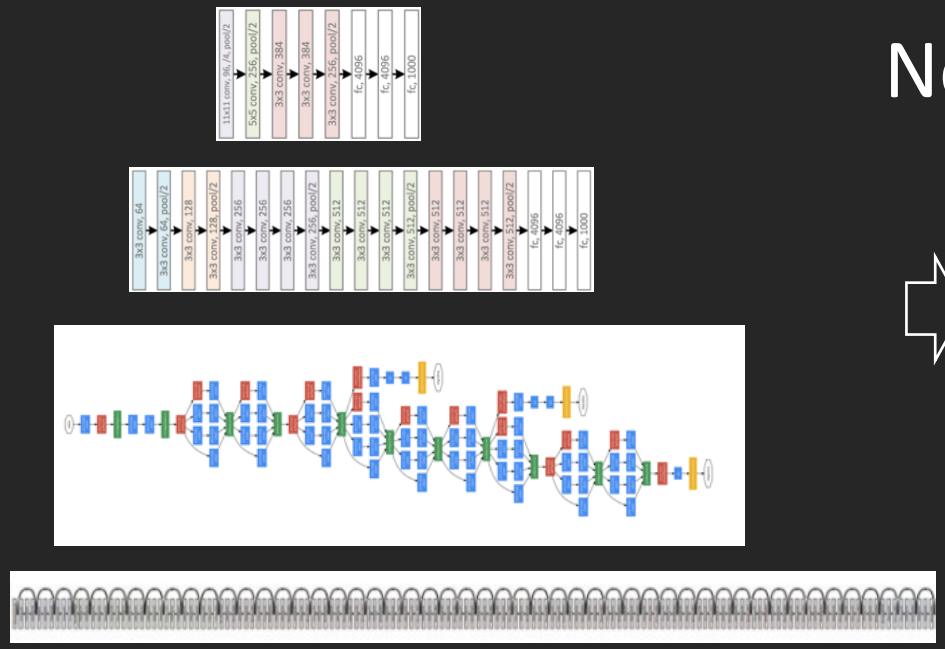
ResNet

Data sources

IMAGENET

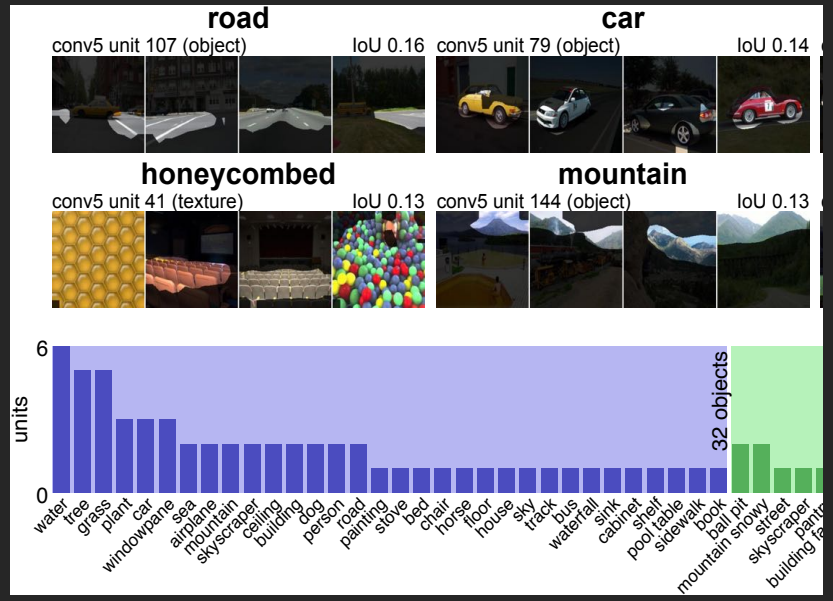
places

Quantify the Interpretability of Networks



Network Dissection

Interpretable Units

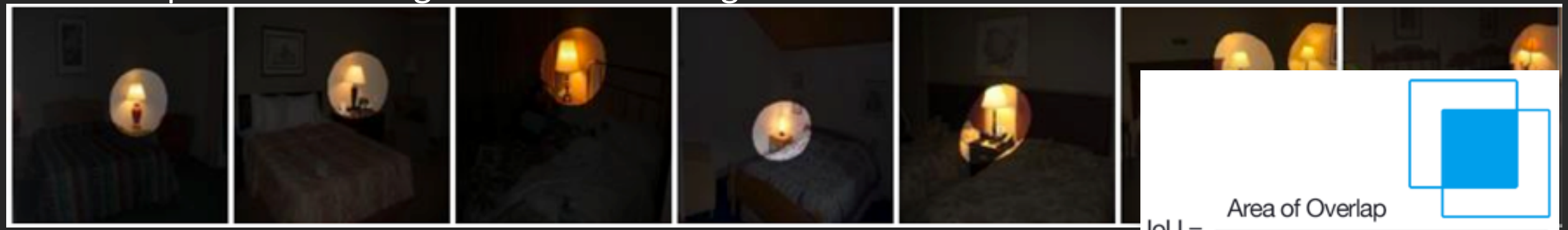


[Bau*, Zhou*, Khosla, Oliva, Torralba. CVPR 2017]

Evaluate Unit for Semantic Segmentation

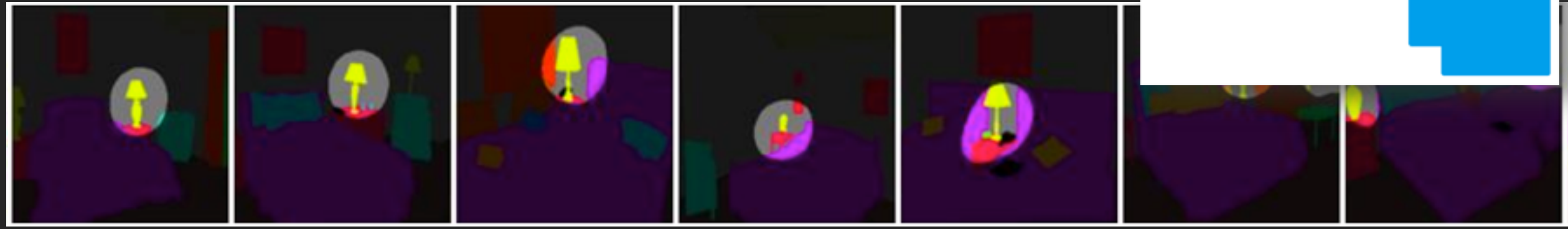
Testing Dataset: 60,000 images annotated with 1,200 concepts

Unit 1: Top activated images from the Testing Dataset



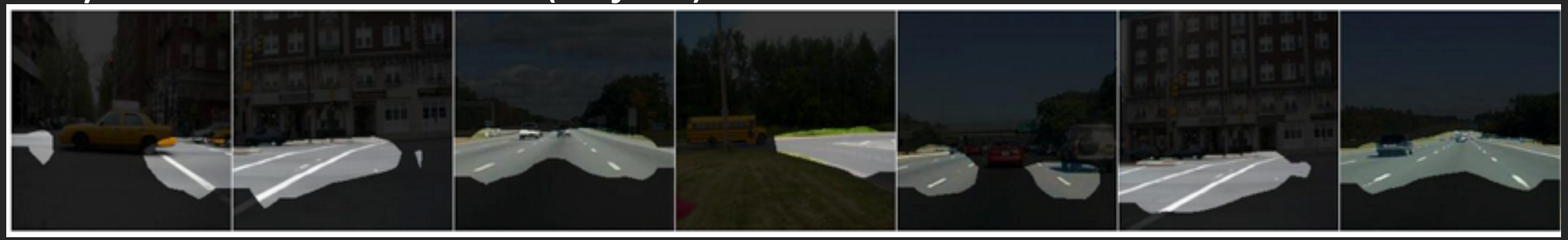
$$IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

Top Concept: Lamp, Intersection over Union (IoU)= 0.23

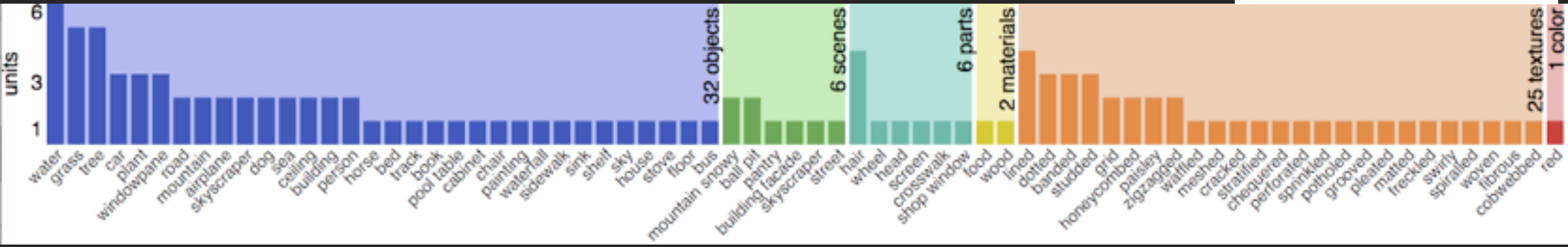


Layer5 unit 79 car (object) IoU=0.13

Layer5 unit 107 road (object) IoU=0.15

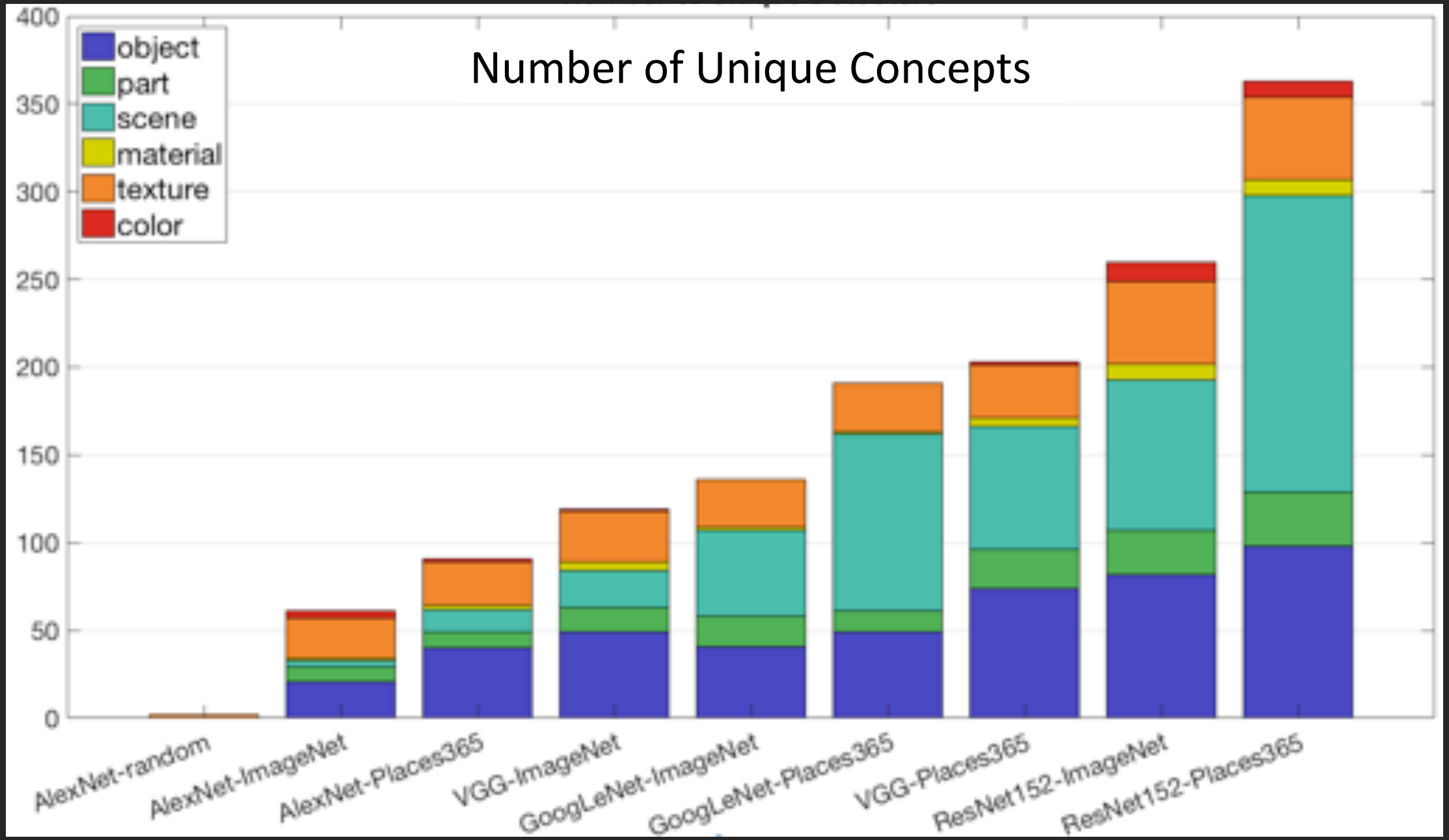


118/256 units covering 72 unique concepts



places
THE SCENE RECOGNITION DATABASE

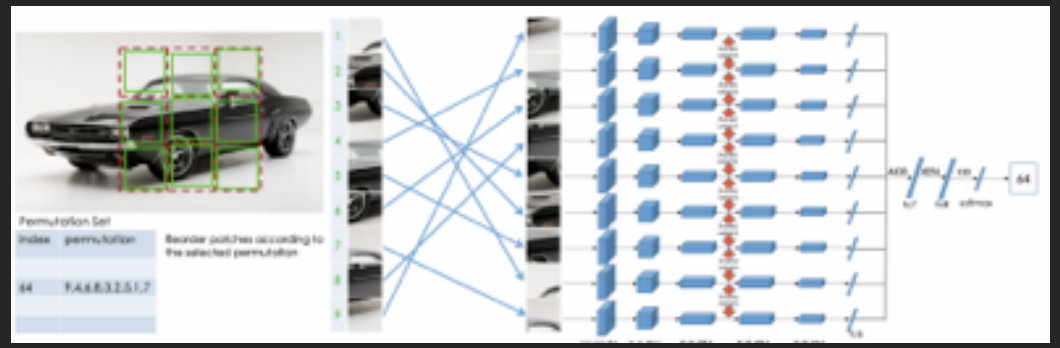

House
Airplane
AlexNet
VGG
GoogLeNet
ResNet
conv5 unit 36
IoU=0.053
conv5 unit 13
IoU=0.101
conv5_3 unit 243
IoU=0.070
conv5_3 unit 151
IoU=0.150
inception_4e unit 789
IoU=0.137
inception_4e unit 92
IoU=0.164
res5c unit 1410
IoU=0.142
res5c unit 1243
IoU=0.172


Number of Unique Concepts
object
part
scene
material
texture
color
400
350
300
250
200
150
100
50
0
AlexNet-random
AlexNet-ImageNet
AlexNet-Places365
VGG-ImageNet
GoogLeNet-ImageNet
GoogLeNet-Places365
VGG-Places365
ResNet152-ImageNet
ResNet152-Places365

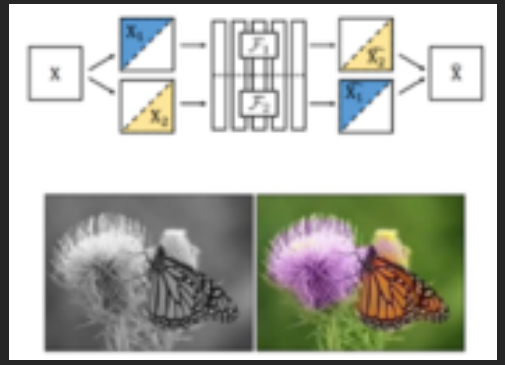
Representations from Self-supervised Learning

Training CNN without image labels.

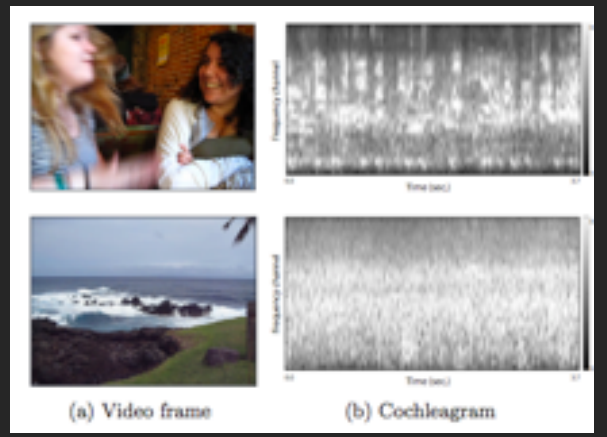
Context prediction, ICCV'15



Solving puzzle, ECCV'16

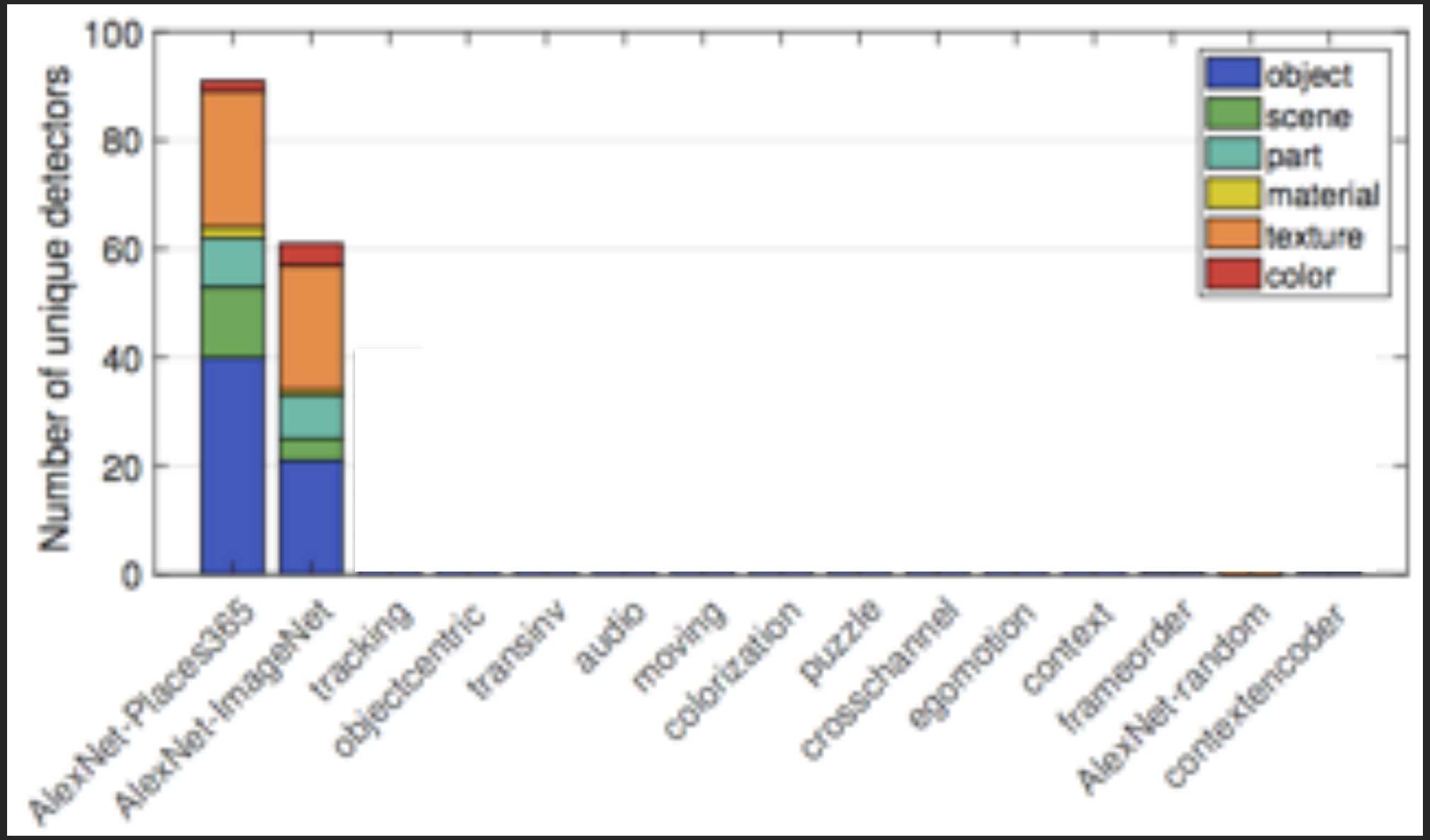


Colorization, ECCV'16 and CVPR'17

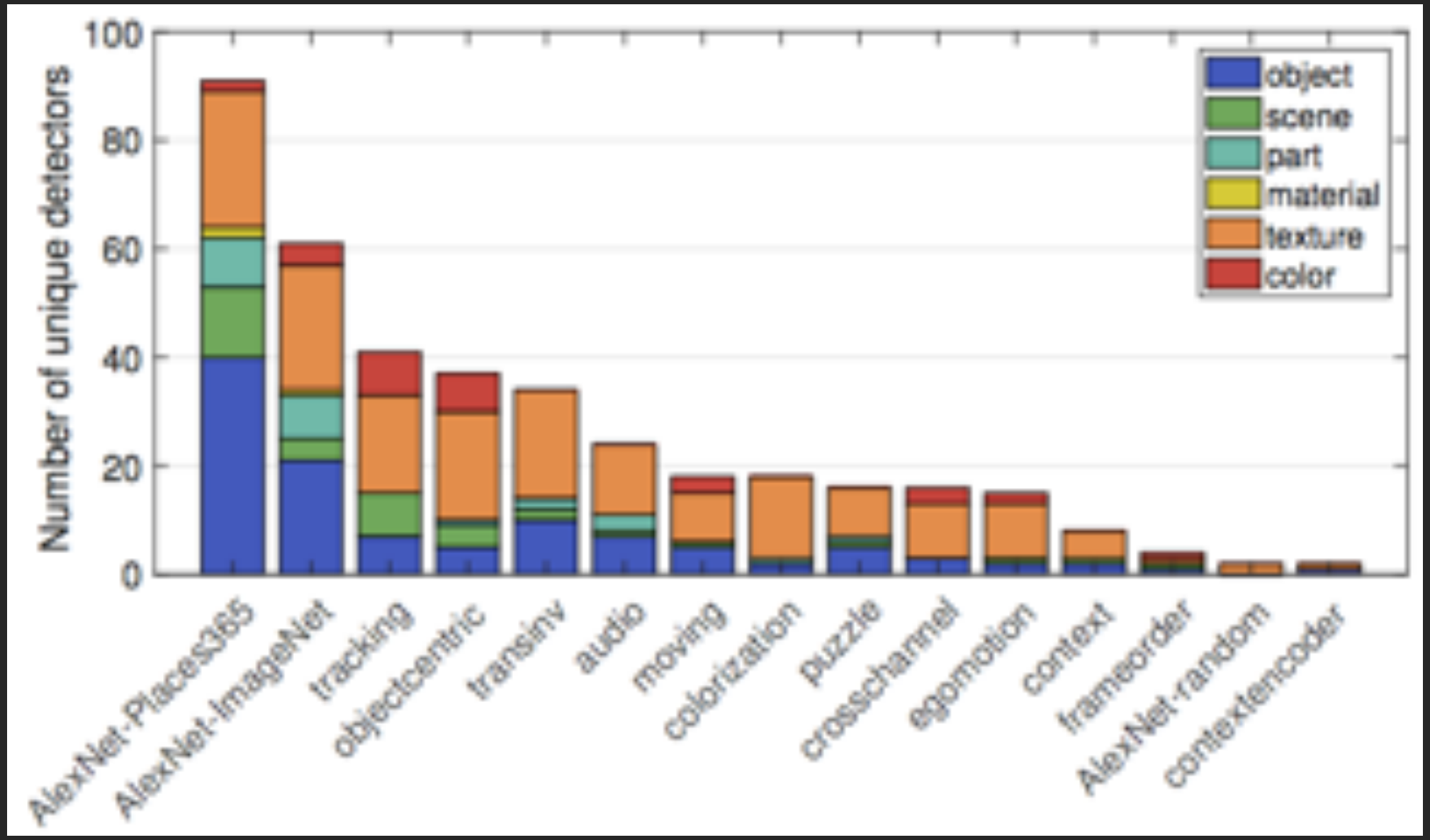


Audio prediction, ECCV'16

Comparison of Supervisions

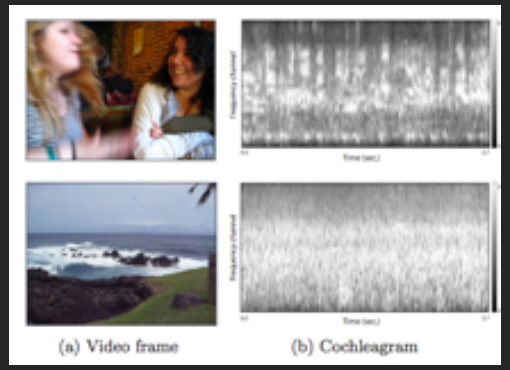


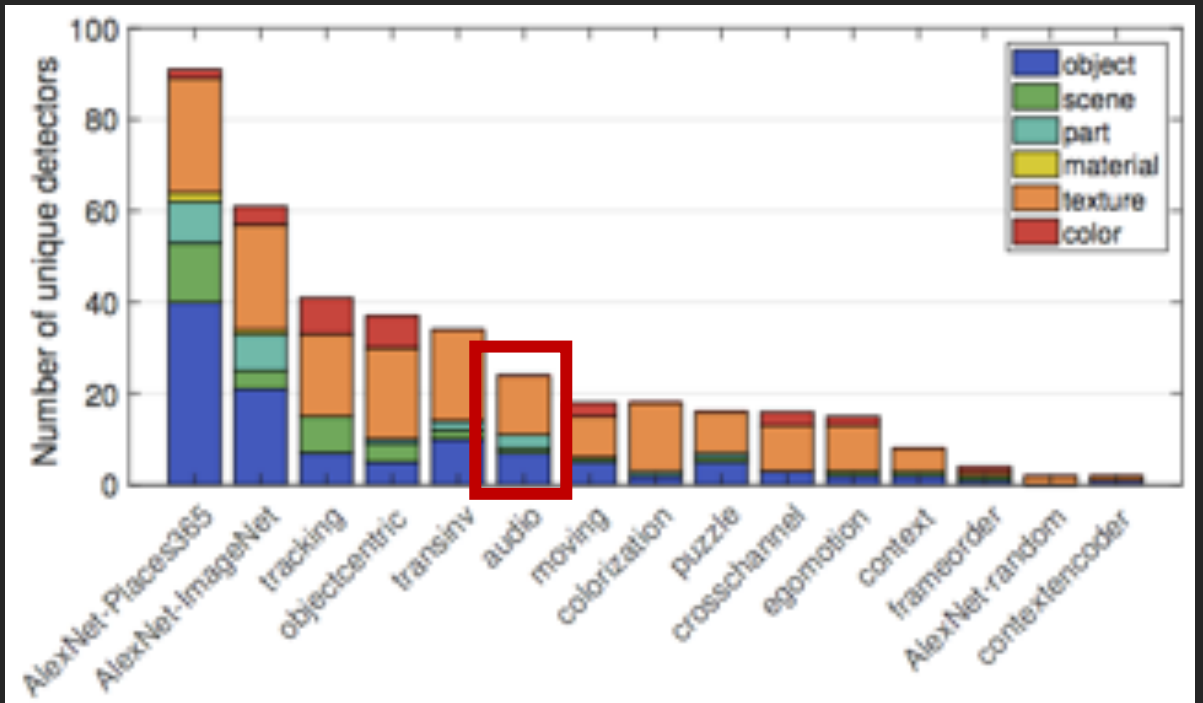
Comparison of Supervisions



Interpretable Units in Self-supervised Networks

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

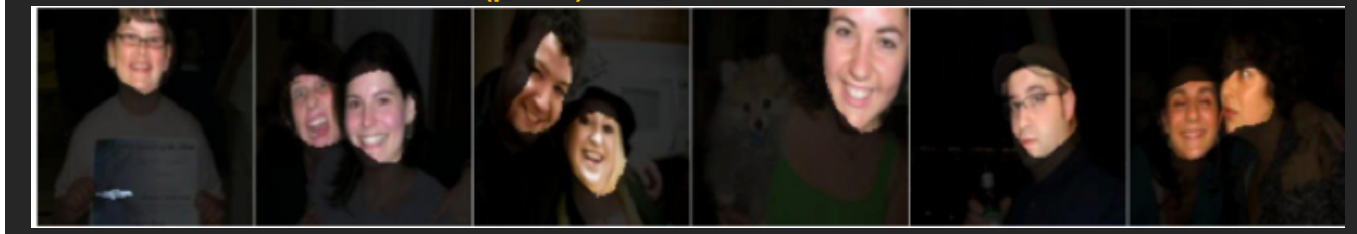




conv5 unit 205: car (object) IoU=0.063

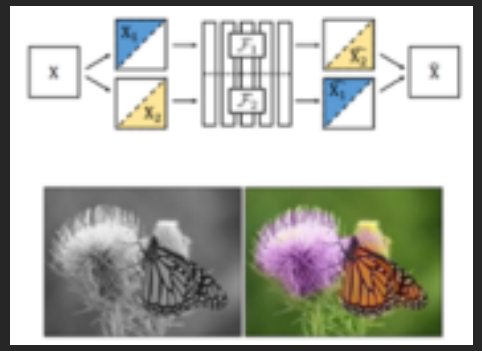
conv5 unit 124: creek (scene) IoU=0.031

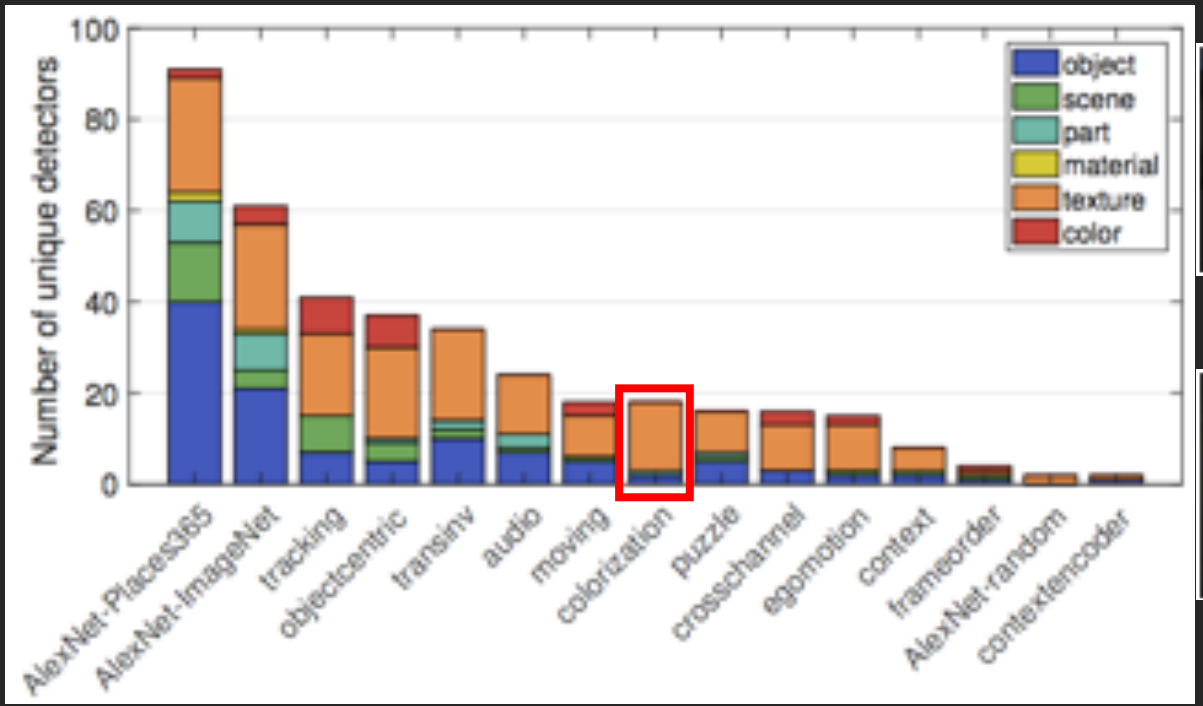
conv5 unit 51: head (part) IoU=0.061



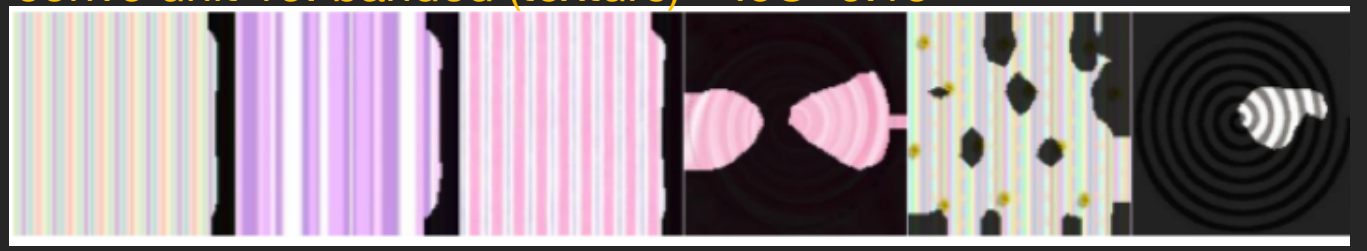
Interpretable Units in Self-supervised Networks

Colorize grey images. ECCV'16. Zhang et al.





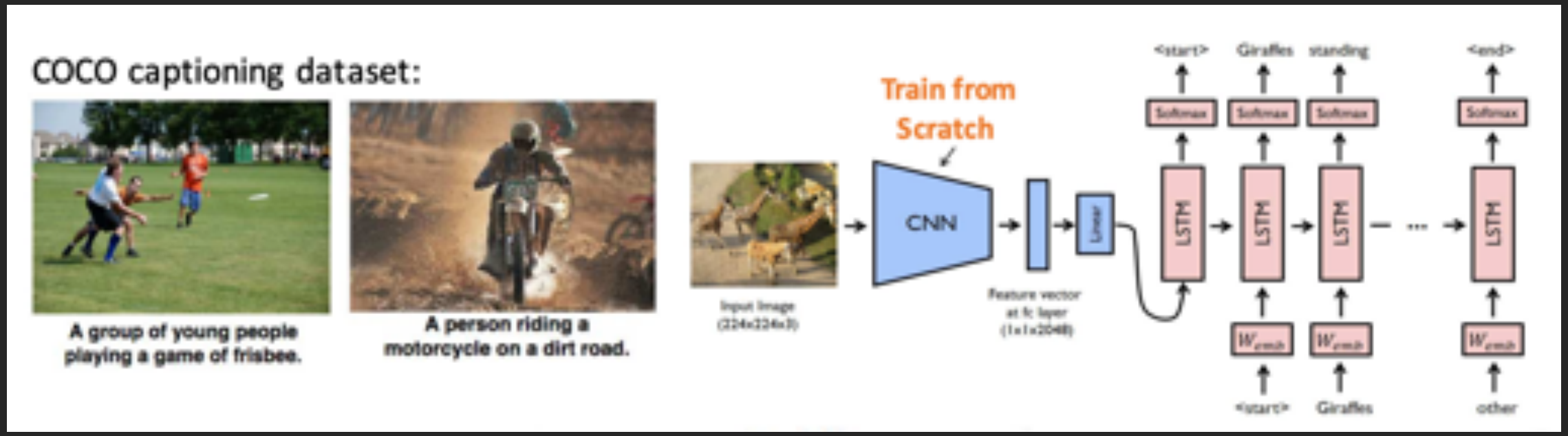
conv5 unit 15: banded (texture) IoU=0.13

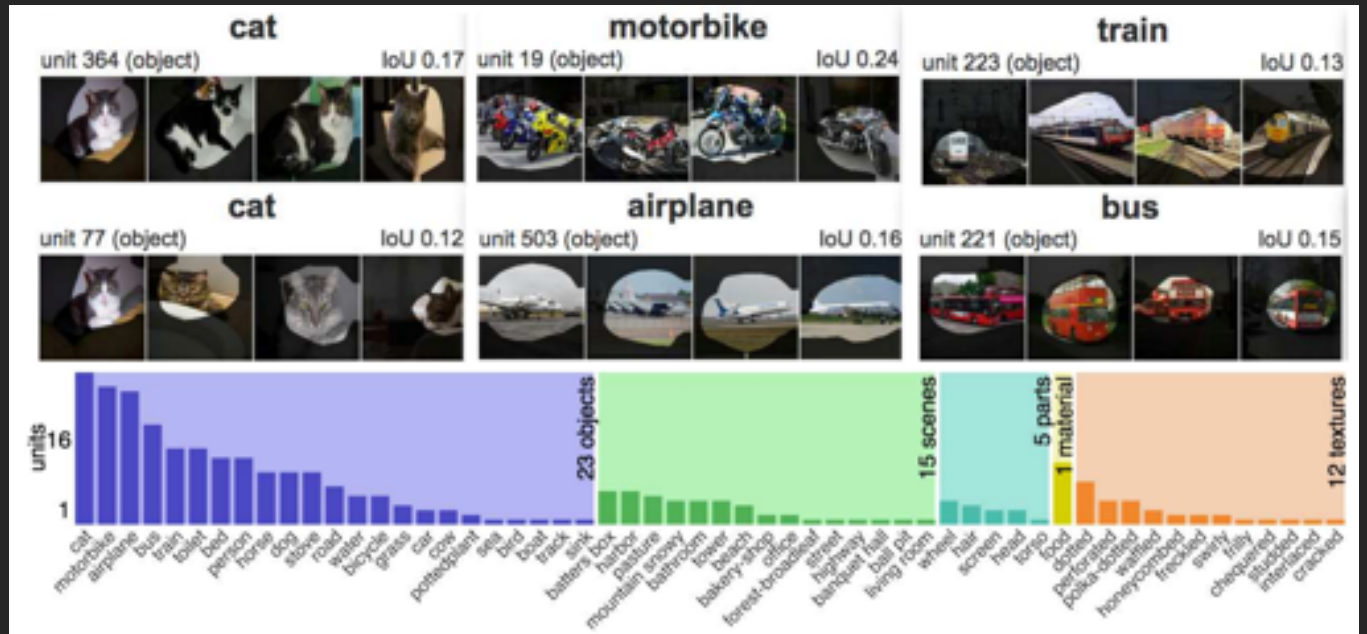


conv5 unit 159: tree (object) IoU=0.039

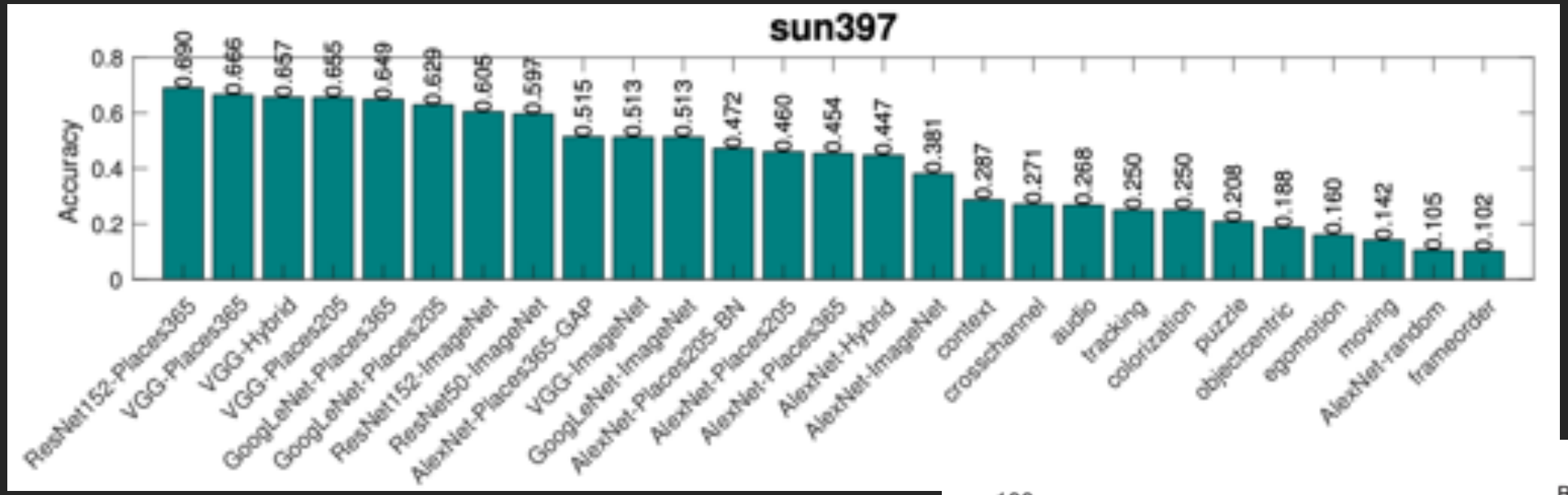
conv5 unit 210: head (part) IoU=0.038

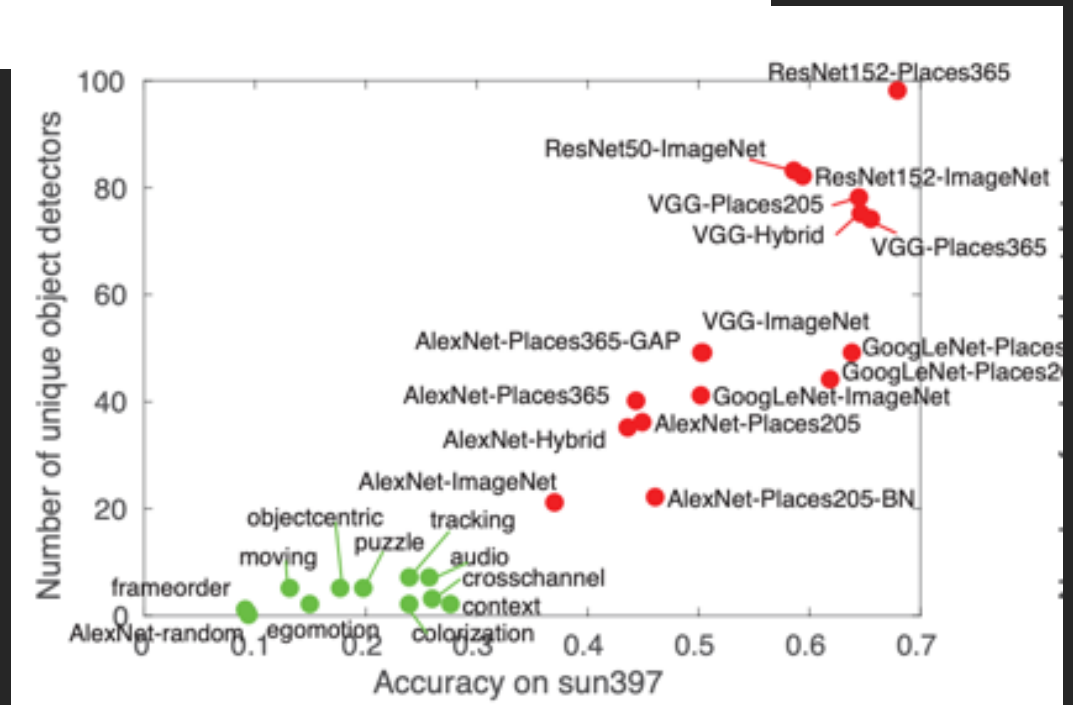
Semantic supervision from image captioning





Classification accuracy for different features





Redefining objective of feature learning

How to design self-supervision tasks for learning discriminative features

How to make features grow more high-level detectors

Rethinking self-supervised feature learning

Issue 1: ImageNet supervised feature is the upper bound???

method	architecture	#params (M)	accuracy (%)
Exemplar [17]	R50w3×	211	46.0 [38]
RelativePosition [13]	R50w2×	94	51.4 [38]
Jigsaw [45]	R50w2×	94	44.6 [38]
Rotation [19]	Rv50w4×	86	55.4 [38]
Colorization [64]	R101*	28	39.6 [14]
DeepCluster [3]	VGG [53]	15	48.4 [4]
BigBiGAN [16]	R50	24	56.6
	Rv50w4×	86	61.3
methods based on contrastive learning follow:			
InstDisc [61]	R50	24	54.0
LocalAgg [66]	R50	24	58.8
CPC v1 [46]	R101*	28	48.7
CPC v2 [35]	$R170^{*}_{wider}$	303	65.9
CMC [56]	$R50_{L+ab}$	47	64.1†
	$R50w2\times_{L+ab}$	188	68.4†
AMDIM [2]	$AMDIM_{small}$	194	63.5†
	$AMDIM_{large}$	626	68.1†
MoCo	R50	24	60.6
	RX50	46	63.9
	R50w2×	94	65.4
	R50w4×	375	**68.6**

Method	ImageNet
ResNet50v2 (sup)	74.4
AMDIM (sup)	71.3

https://arxiv.org/pdf/1911.05722.pdf

Rethinking self-supervised feature learning

Can we treat ImageNet feature as lower bound?

Google's JET-300M data

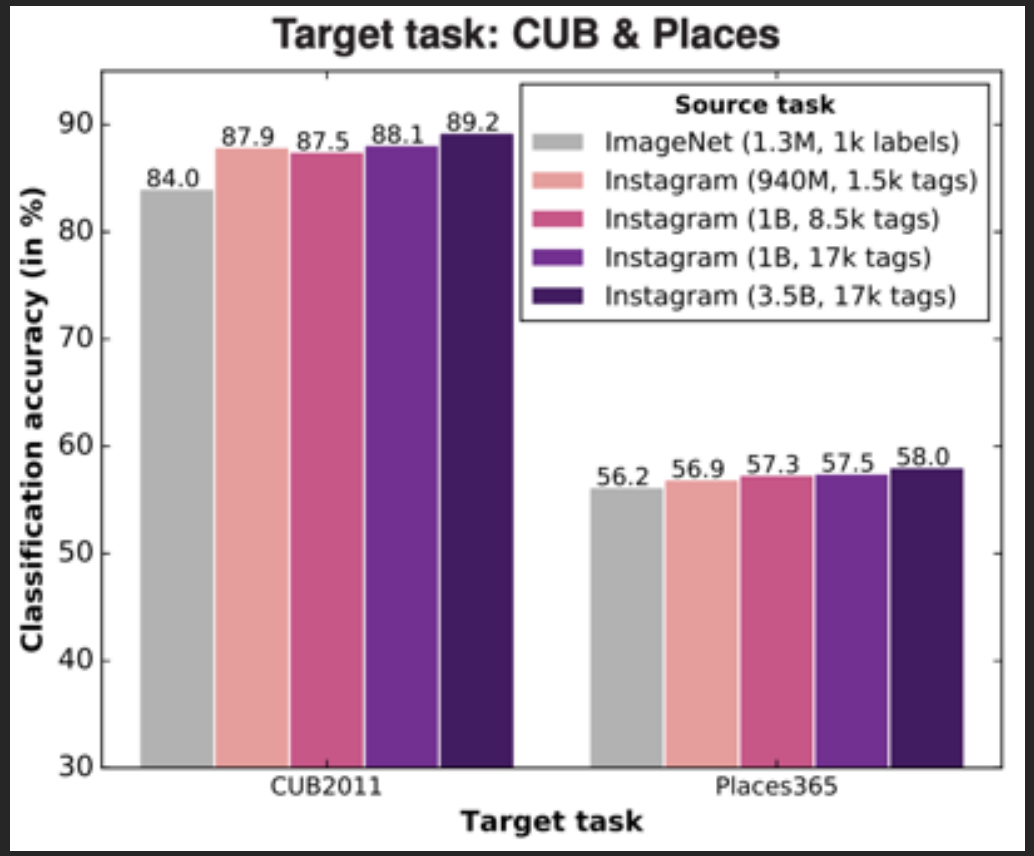
Initialization	Top-1 Acc.	Top-5 Acc.
MSRA checkpoint [16]	76.4	92.9
Random initialization	77.5	93.9
Fine-tune from JFT-300M	**79.2**	**94.7**

Table 1. Top-1 and top-5 classification accuracy on the ImageNet 'val' set (single model and single crop inference are used).

Method	mAP@0.5	mAP@[0.5,0.95]
He *et al.* [16]	53.3	32.2
ImageNet	53.6	34.3
300M	56.9	36.7
ImageNet+300M	**58.0**	**37.4**
Inception ResNet [38]	56.3	35.5

Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. Chen Sun et al. ICCV'17. Google

Facebook's Instagram hashtag 1B



Exploring the Limits of Weakly Supervised Pretraining. D.Hahajan et al. ECCV'18.

Rethinking self-supervised feature learning

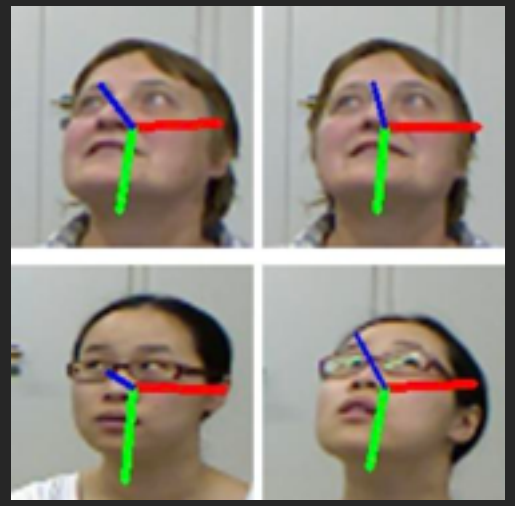
Issue 2: Self-supervision = get a feature for ImageNet classification?

Method	ImageNet
ResNet50v2 (sup)	74.4
AMDIM (sup)	71.3

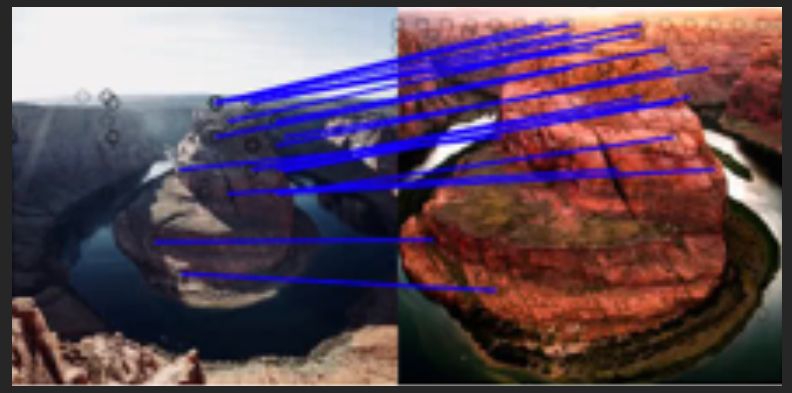
method	architecture	#params (M)	accuracy (%)
Exemplar [17]	R50w3×	211	46.0 [38]
RelativePosition [13]	R50w2×	94	51.4 [38]
Jigsaw [45]	R50w2×	94	44.6 [38]
Rotation [19]	Rv50w4×	86	55.4 [38]
Colorization [64]	R101*	28	39.6 [14]
DeepCluster [3]	VGG [53]	15	48.4 [4]
BigBiGAN [16]	R50	24	56.6
	Rv50w4×	86	61.3
methods based on contrastive learning follow:			
InstDisc [61]	R50	24	54.0
LocalAgg [66]	R50	24	58.8
CPC v1 [46]	R101*	28	48.7
CPC v2 [35]	$R170^{*}_{wider}$	303	65.9
CMC [56]	$R50_{L+ab}$	47	64.1†
	$R50w2\times_{L+ab}$	188	68.4†
AMDIM [2]	$AMDIM_{small}$	194	63.5†
	$AMDIM_{large}$	626	68.1†
MoCo	R50	24	60.6
	RX50	46	63.9
	R50w2×	94	65.4
	R50w4×	375	**68.6**

Other tasks beyond classification

Pose estimation

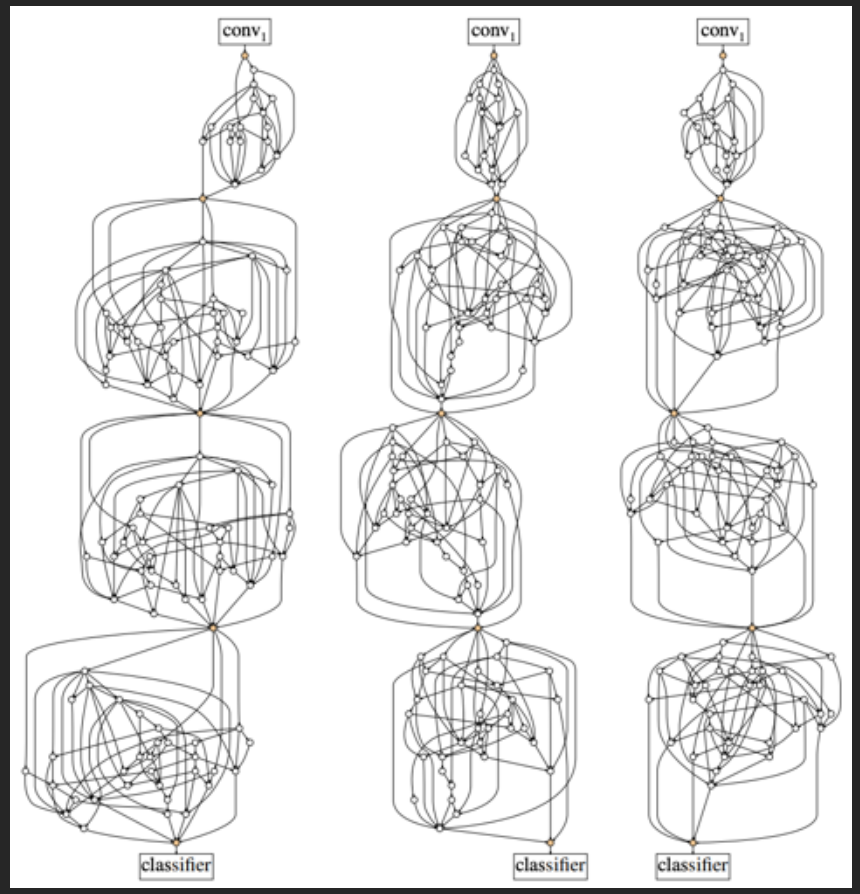


Keypoint alignment

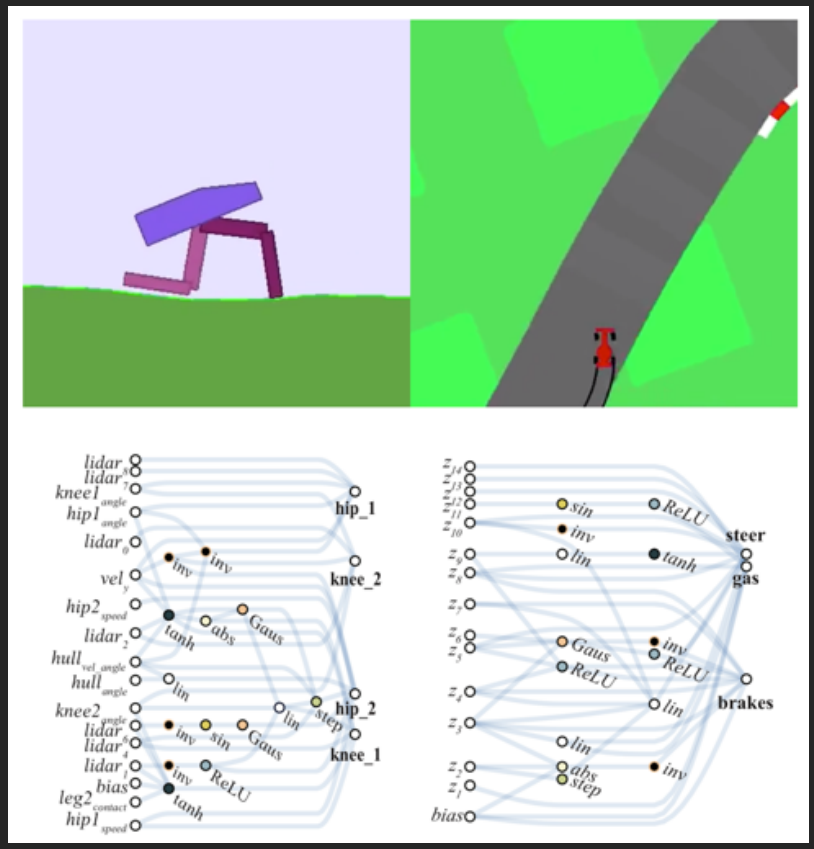


Rethinking self-supervised feature learning

Issue 3: Self-supervision/supervision = just training weights for AlexNet/ResNet?



Exploring Randomly Wired Neural Networks for Image Recognition. Saining Xie, et al. ICCV'19



Weight Agnostic Neural Networks. Adam Gaier, David Ha. NeurIPS'19. https://weightagnostic.github.io/