Network Dissection: Quantifying Interpretability of Deep Visual Representation

Bolei Zhou*

MIT

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

Seminar/SS18: Explainable Machine Learning

21.06.2018

Presenter: Pingchuan Ma

Agenda

1. Motivation

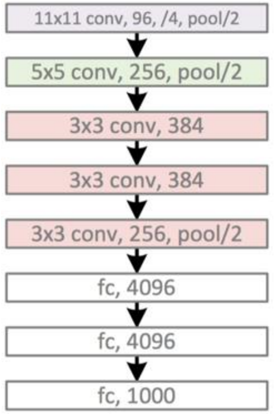
2. Definition, Dataset and Method

3. Experiments

4. Conclusion

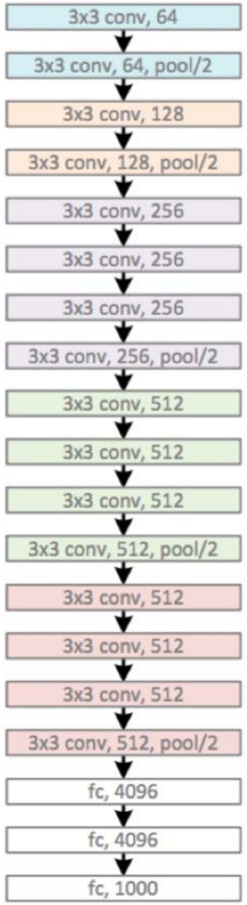
Motivation – Why we study interpretable units?

2012: AlexNet
5 conv. layers



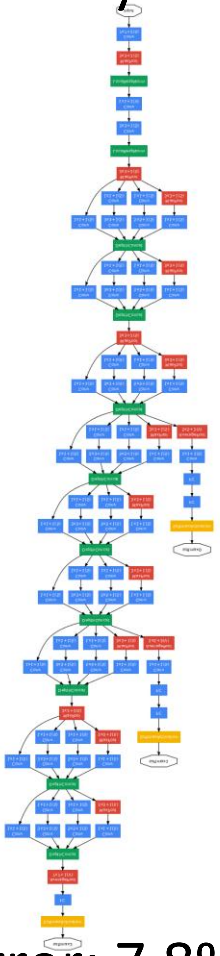
Error: 15.3%

2014: VGG
16 conv. layers



Error: 8.5%

2015: GoogLeNet
22 conv. layers



Error: 7.8%

2016: ResNet
>100 conv. layers

Error: 4.4%

Motivation – Why we study interpretable units?

1. High performance but black boxes lack interpretability

2. Human want to understand things, especially those tools that we count on

3. Interpretable units hint that deep network may not be completely black boxes

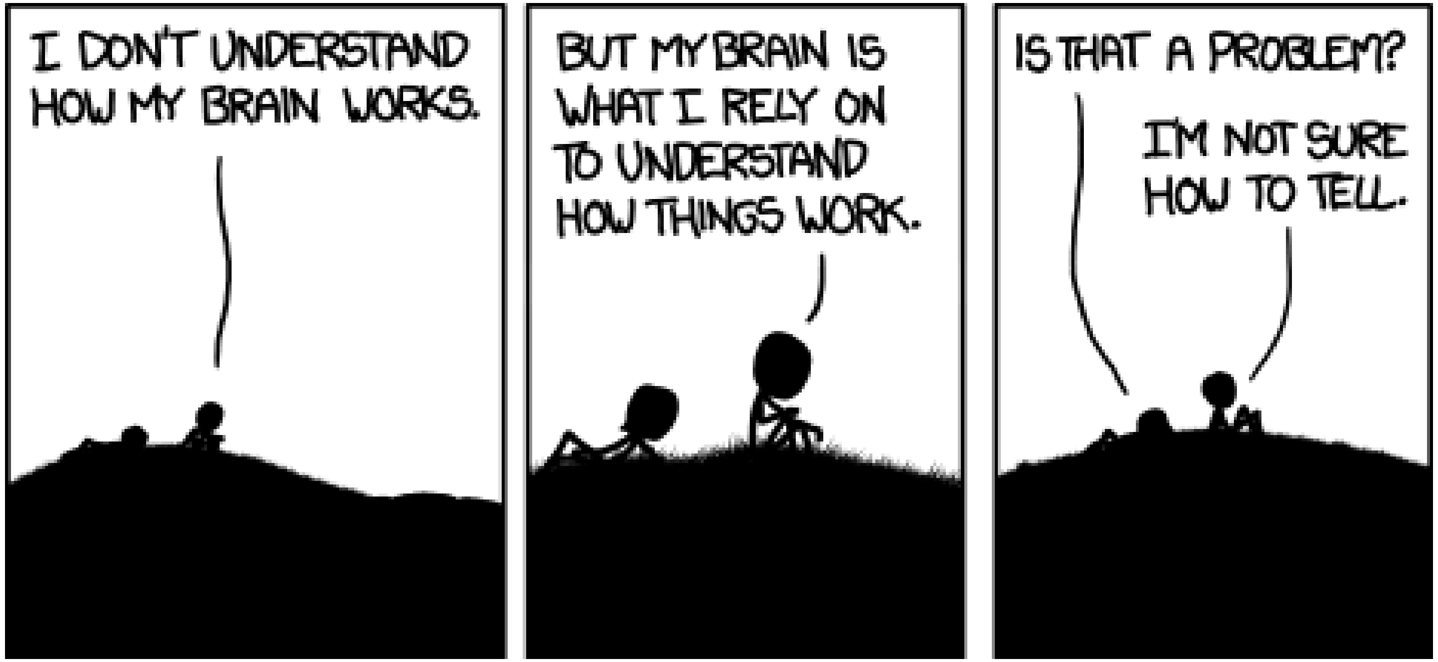
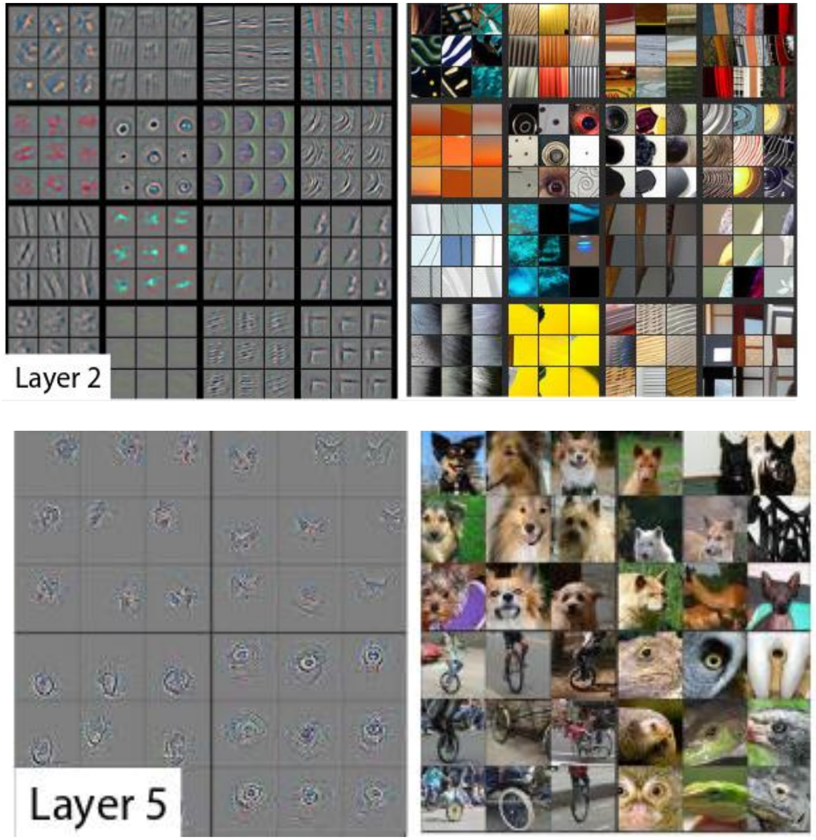


Fig.1
by Matt Scherer

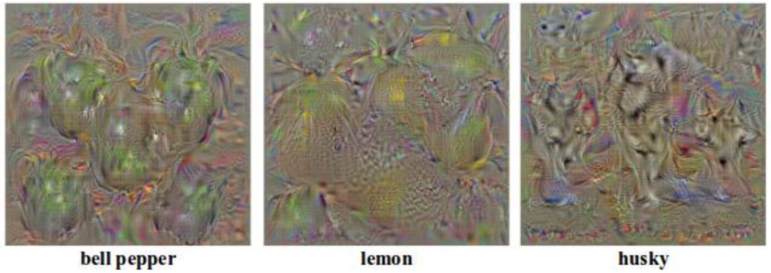
Motivation – Previous and related work

Deconvolution



Zeiler et al., ECCV 2014.

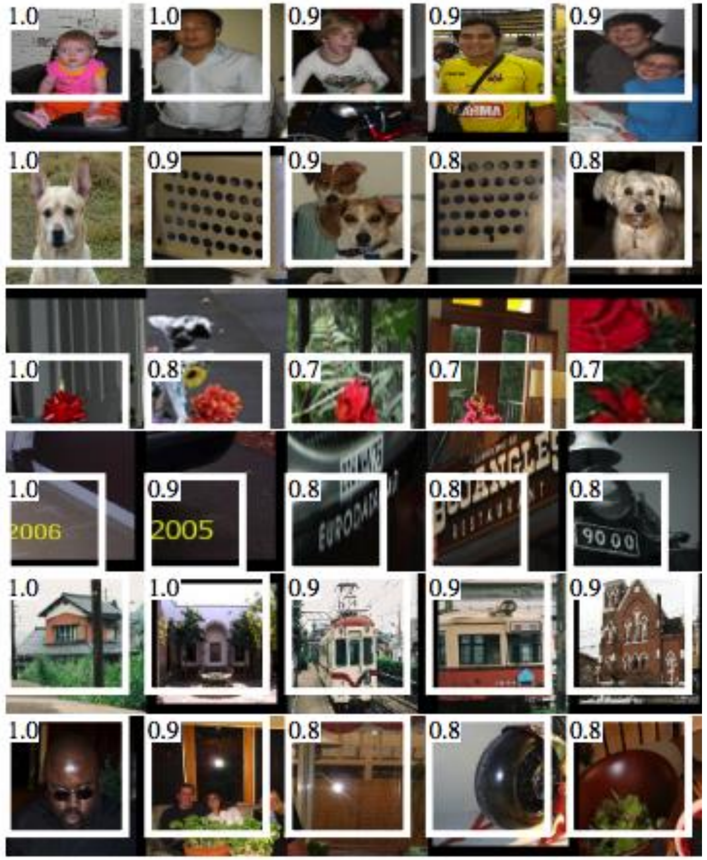
Back-propagation



Simonyan et al., ICLR 2015

Inceptionism. Google Blog. June 2015

Top activated images



Girshick et al., CVPR 2014

Definition – Disentangled representation

1. CNNs may be learning spontaneously the *disentangled representation*, **which aligns its variables with a meaningful factorization of the underlying problem structure.**

2. Partly disentangled for economical use of hidden variables.

3. To detect those disentangled structure and simply read out the separated factors

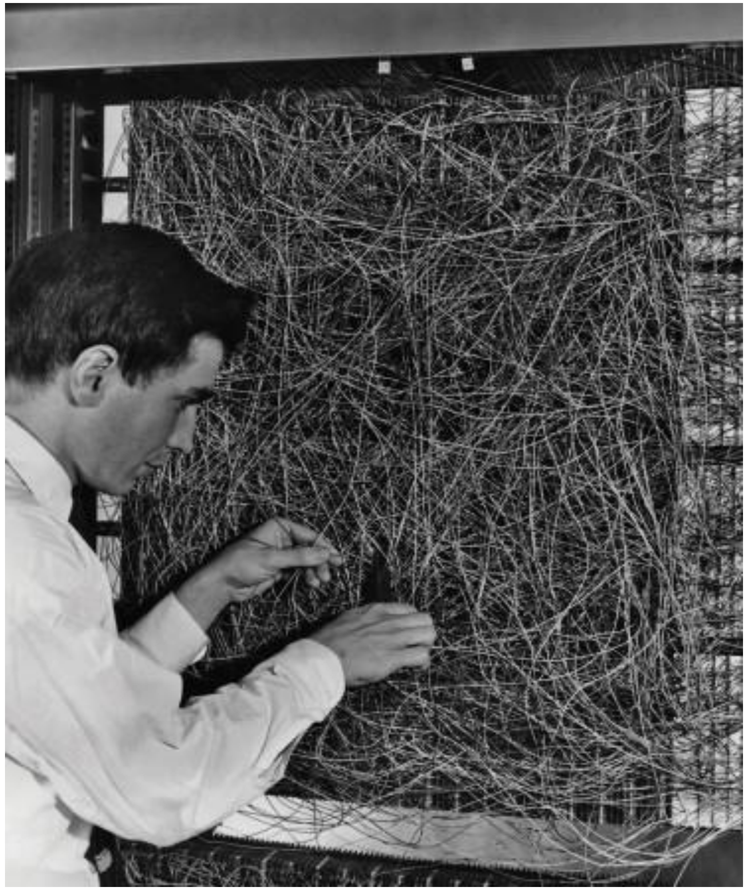
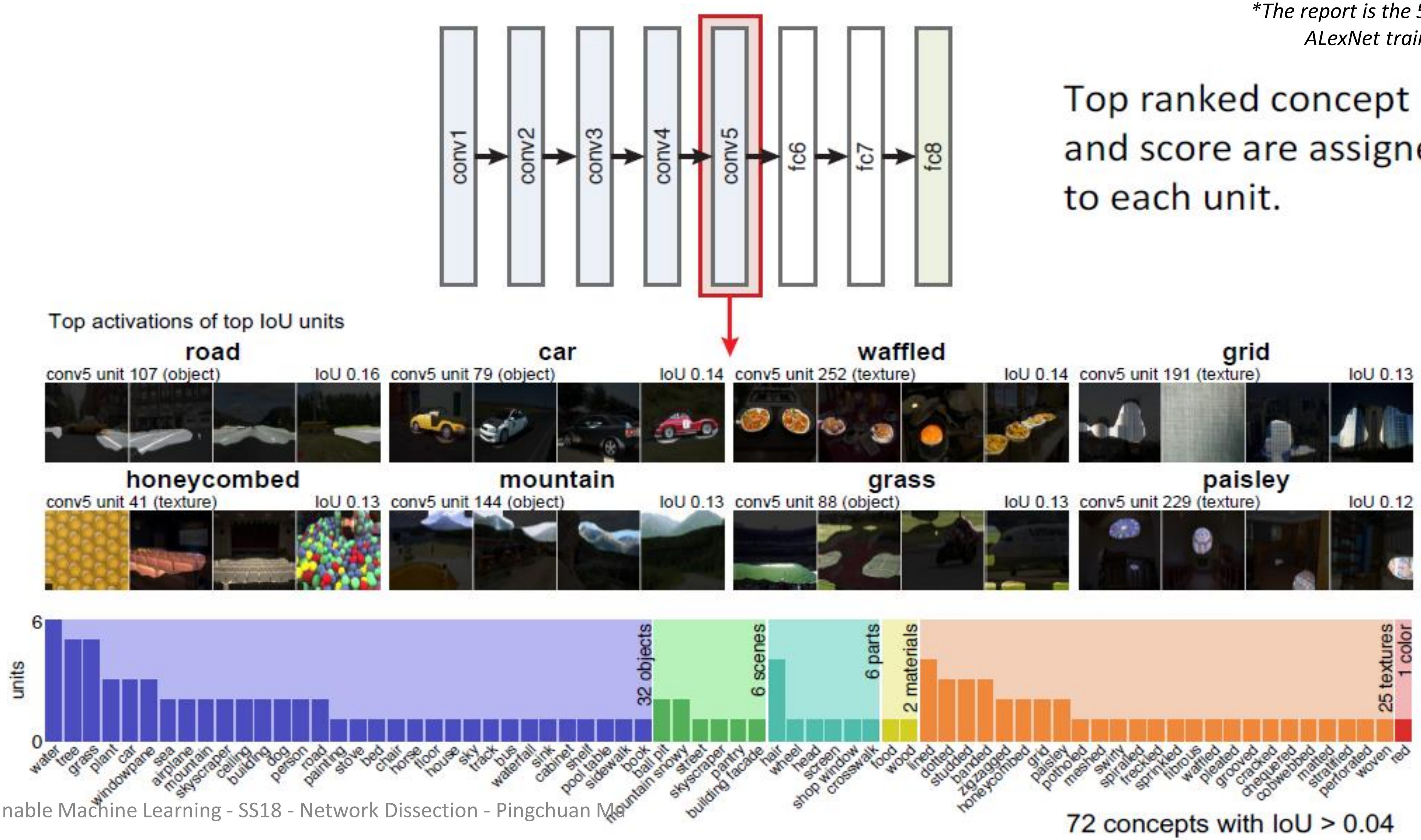


Fig.2
Early artificial neural network,
at the Cornell Aeronautical Laboratory
in Buffalo, New York

Definition – Network Dissection, a tool kit

**The report is the 5 conv layer from a ALexNet trained on Places*



Top ranked concept and score are assigned to each unit.

Definition – Steps to Quantify Interpretability

Step 1. Identify a broad set of human-labeled visual concepts. *(Broden Dataset)*

Step 2. Gather hidden variables' response to known concepts. *(Distribution of individual unit activation beyond a certain threshold)*

Step 3. Quantify alignment of hidden variable-concept pairs. *(Calculate the IoU of them) Single hidden units in network and single concepts in Broden*

Step 1: Dataset – Broden

Not misspell

***Bro**adly and **Den**sely Labeled Dataset*, namely Broden, unifies several densely labeled datasets.

Purpose: to provide a ground truth set of exemplars of visual concepts, which are normalized and cleaned.

Total = 63,305 images
1,197 visual concepts

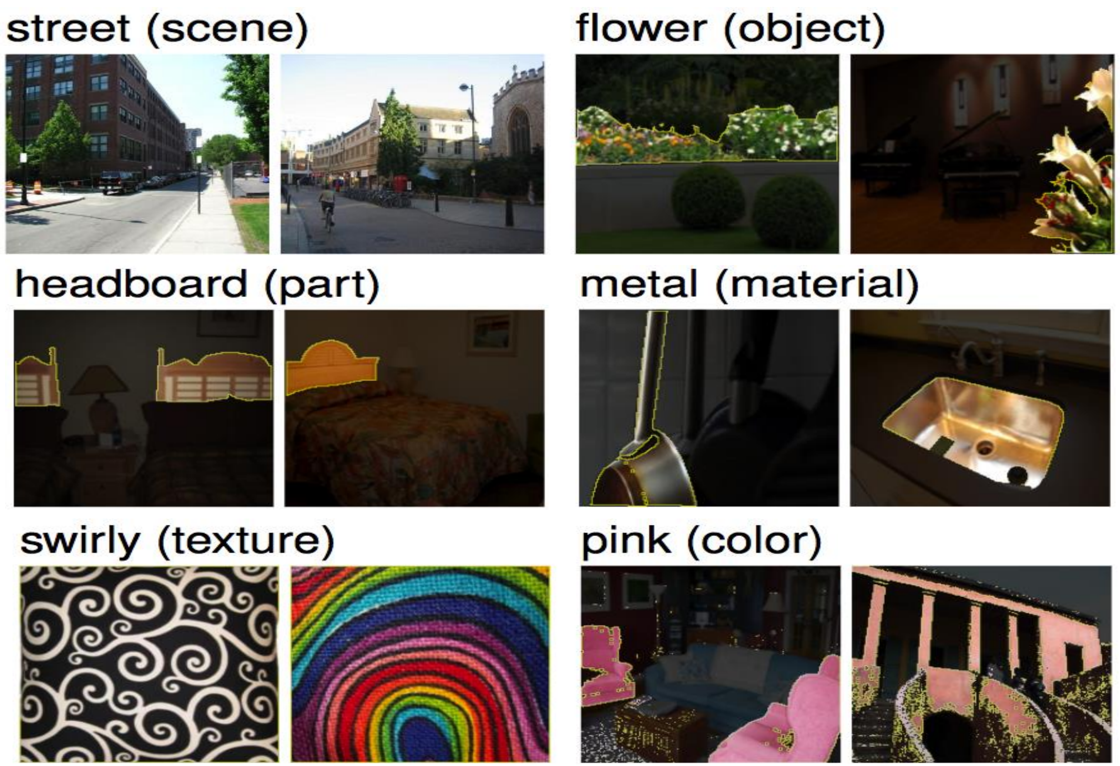
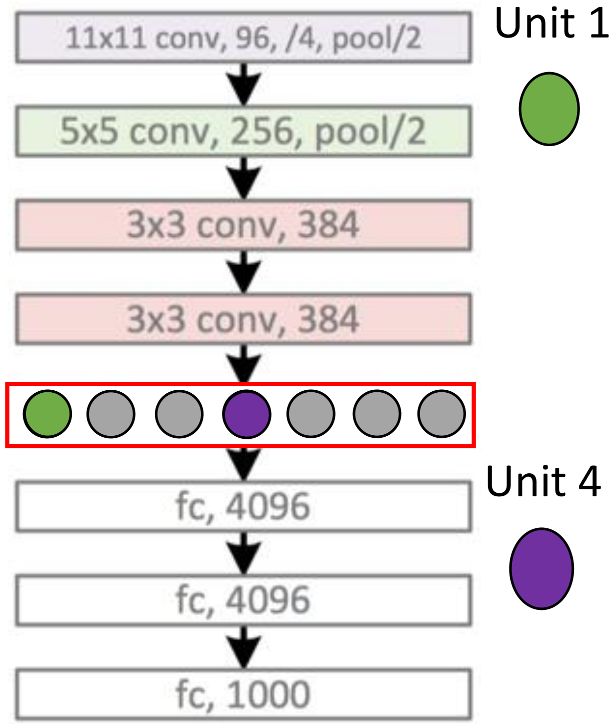


Table 1. Statistics of each label type included in the data set.

Category	Classes	Sources	Avg sample
scene	468	ADE [43]	38
object	584	ADE [43], Pascal-Context [19]	491
part	234	ADE [43], Pascal-Part [6]	854
material	32	OpenSurfaces [4]	1,703
texture	47	DTD [7]	140
color	11	Generated	59,250

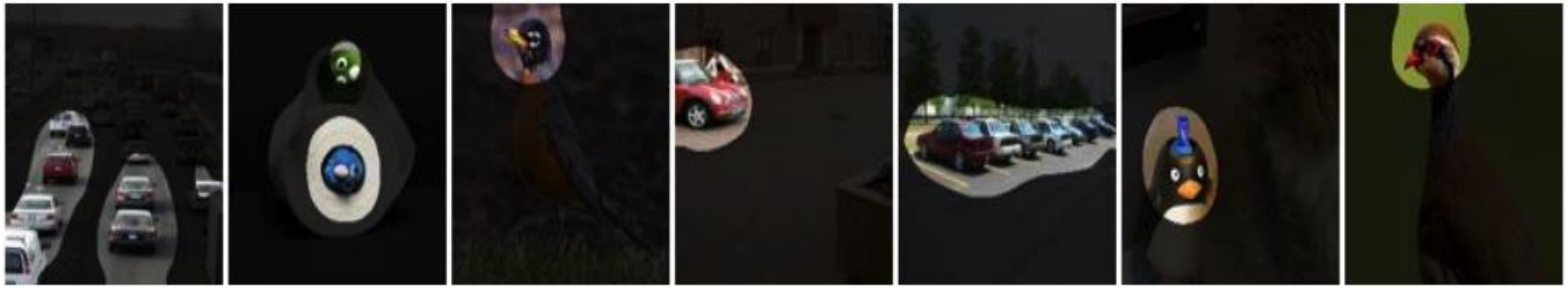
Step 2: Method – Distribution of Activation

Pre-trained model



Top Activated Images *Interpretation: lamp *Score: 0.15

Top Activated Images *Interpretation: car *Score: 0.02

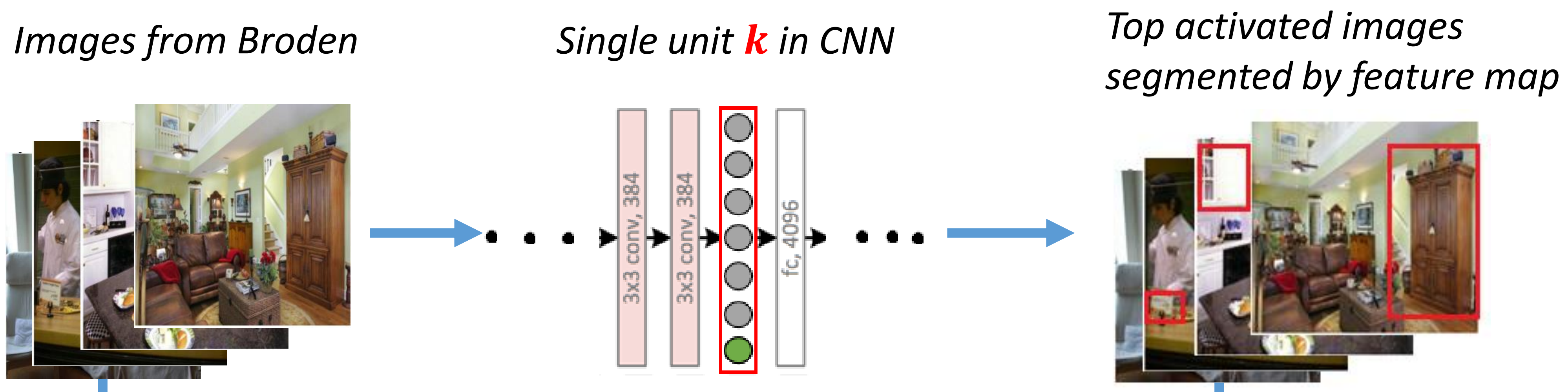


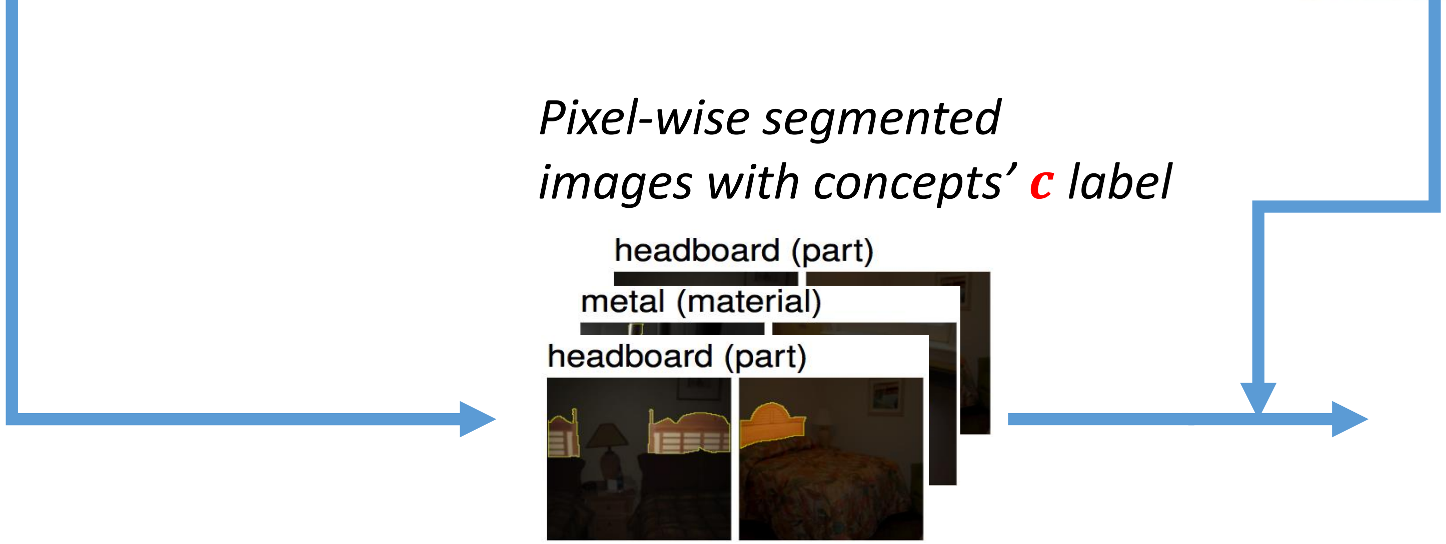
Step 3: Method – IoU

Unit 1 Top activated images

Lamp Intersection over Union (IoU)= 0.12

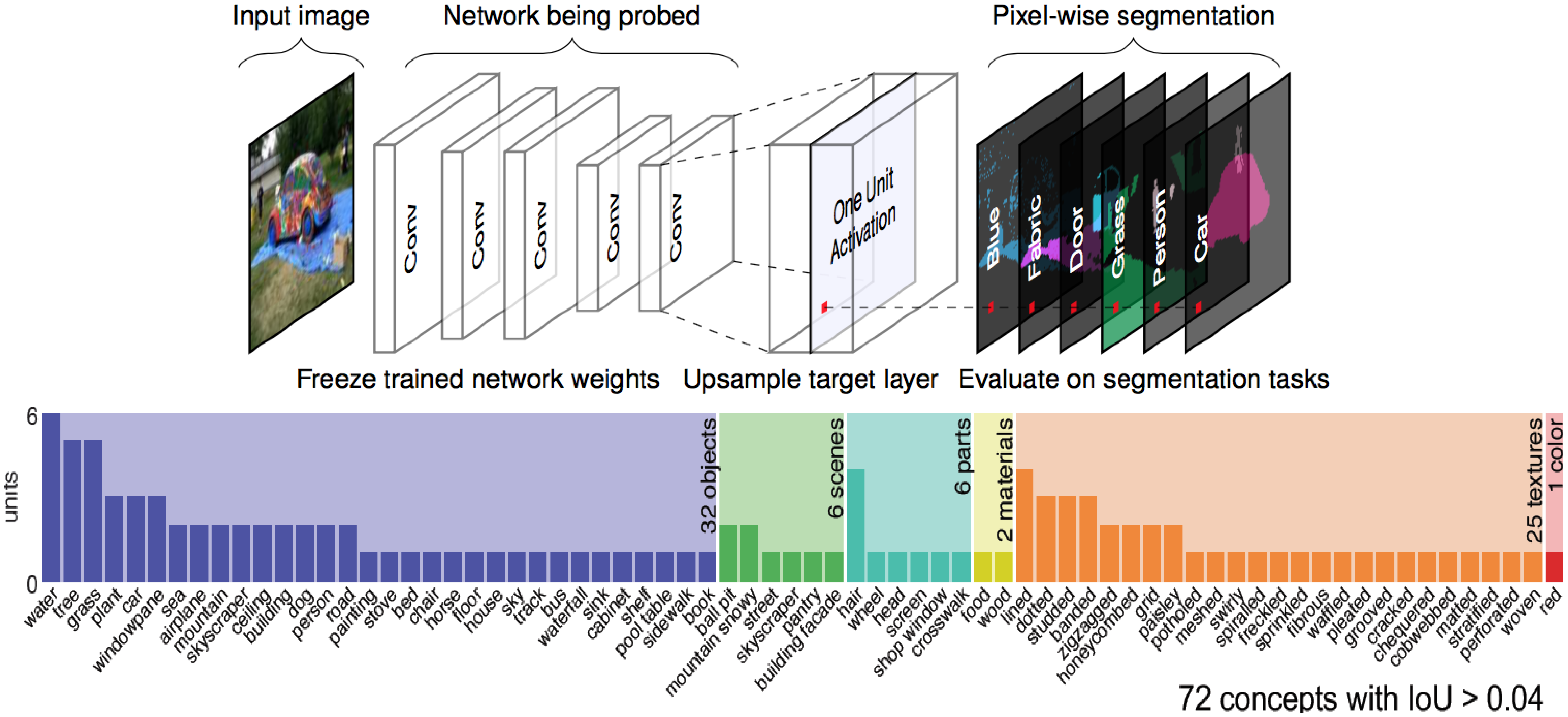
Method – Scoring Unit Interpretability



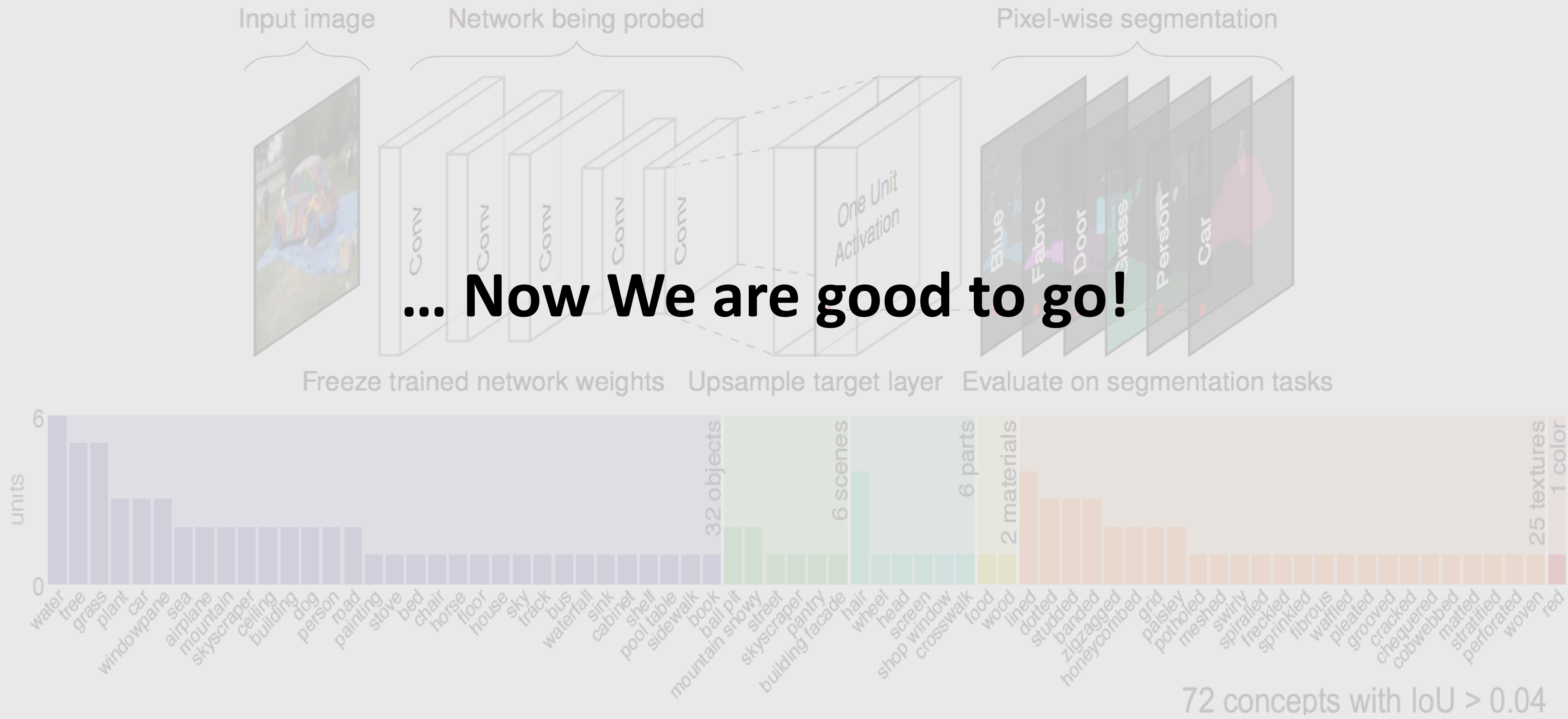


1. Calculate the $IoU_{k,c}$ data-set-wide for every pair of $(\boldsymbol{k}, \boldsymbol{c})$
2. If $IoU_{k,c}$ exceeds a threshold, we consider unit ***k*** as a concept ***c*** detector.

Experiments – Recap



Experiments – Recap



… Now We are good to go!

Experiments – Structure

Steps:

1. Human evaluation
2. **Axis-independent**
3. Layer levels
4. Architectures and supervisions
5. Training conditions
6. **Discrimination vs. Interpretability**
7. **Layer Width vs. Interpretability**
8. **Fine-tuning**

Training	Network	Data set or task
none	AlexNet	random
Supervised	AlexNet	ImageNet, Places205, Places365, Hybrid.
	GoogLeNet	ImageNet, Places205, Places365.
	VGG-16	ImageNet, Places205, Places365, Hybrid.
	ResNet-152	ImageNet, Places365.
Self	AlexNet	context, puzzle, egomotion, tracking, moving, videoorder, audio, crosschannel,colorization. objectcentric.

***Baseline Model:** AlexNet trained on Places205

Experiments – 1. Human evaluation

Evaluation: Amazon Mechanical Turk (AMT)

Method: Rater are shown images patches and are asked yes/no

	conv1	conv2	conv3	conv4	conv5
Interpretable units	57/96	126/256	247/384	258/384	194/256
Human consistency	82%	76%	83%	82%	91%
Network Dissection	37%	56%	54%	59%	71%

Experiments – 2. Axis-independent

Two Hypothesis:

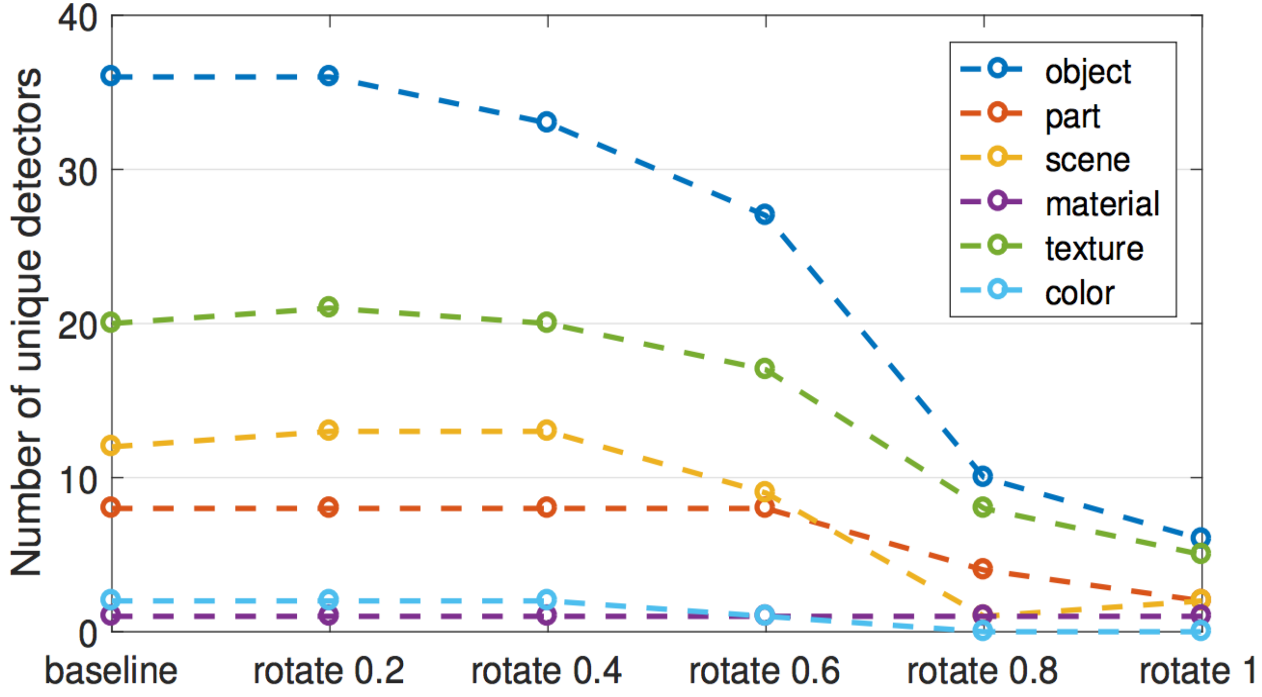
1. The overall level of interpretability should not be affected by a change in rotation.

2. The overall level of interpretability is expected to drop under this change.

Method:

Apply random changes Q in basis to a representation $f(x)$ learned by AlexNet, compare unique detectors

Unique detectors in conv5 layer of Baseline AlexNet



Unique detectors in $\boldsymbol{Qf(x)}$ is much fewer than in $\boldsymbol{f(x)}$

However each rotated representation has exactly the same discriminative power as the original one.

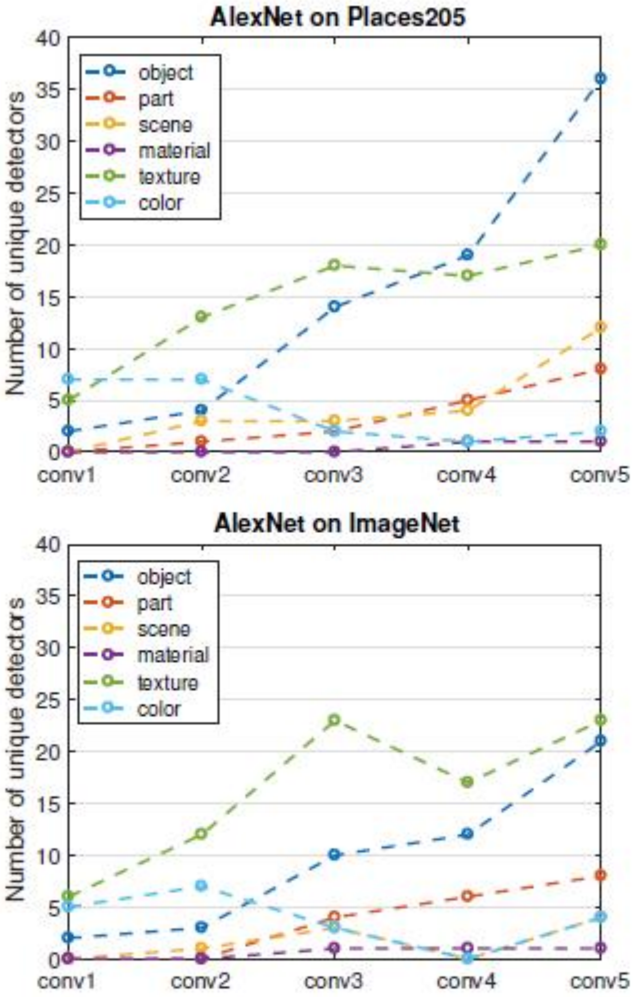
Experiments – 2. Axis-independent

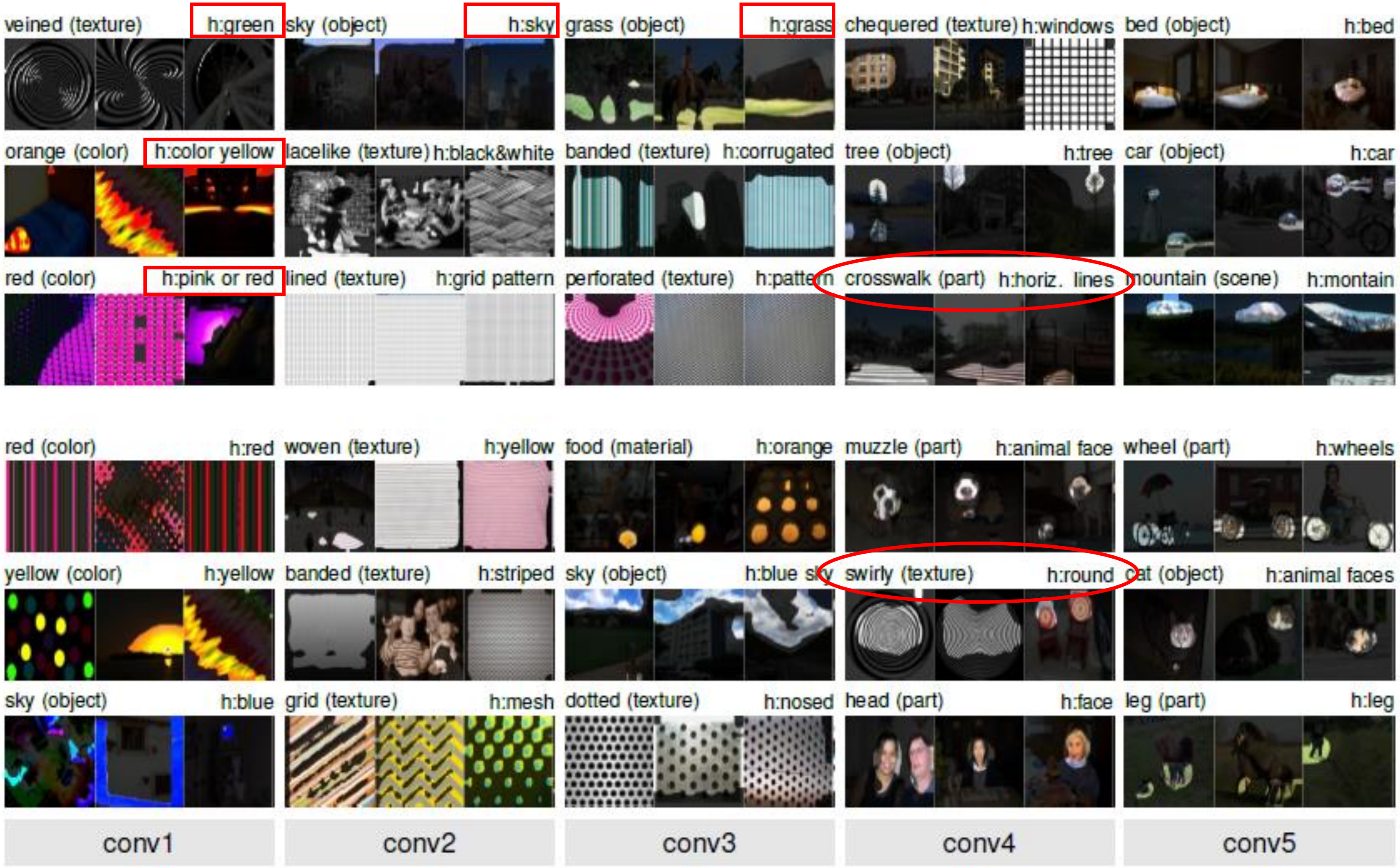
Conclusion:

The interpretability of CNNs is not an axis-independent property, and it is neither an inevitable/ necessary result of the discriminative power of a representation, nor is a prerequisite to discriminative power.

Instead, the interpretability is more likely to be a different quality from discriminative power that must be measured separately to be understood.

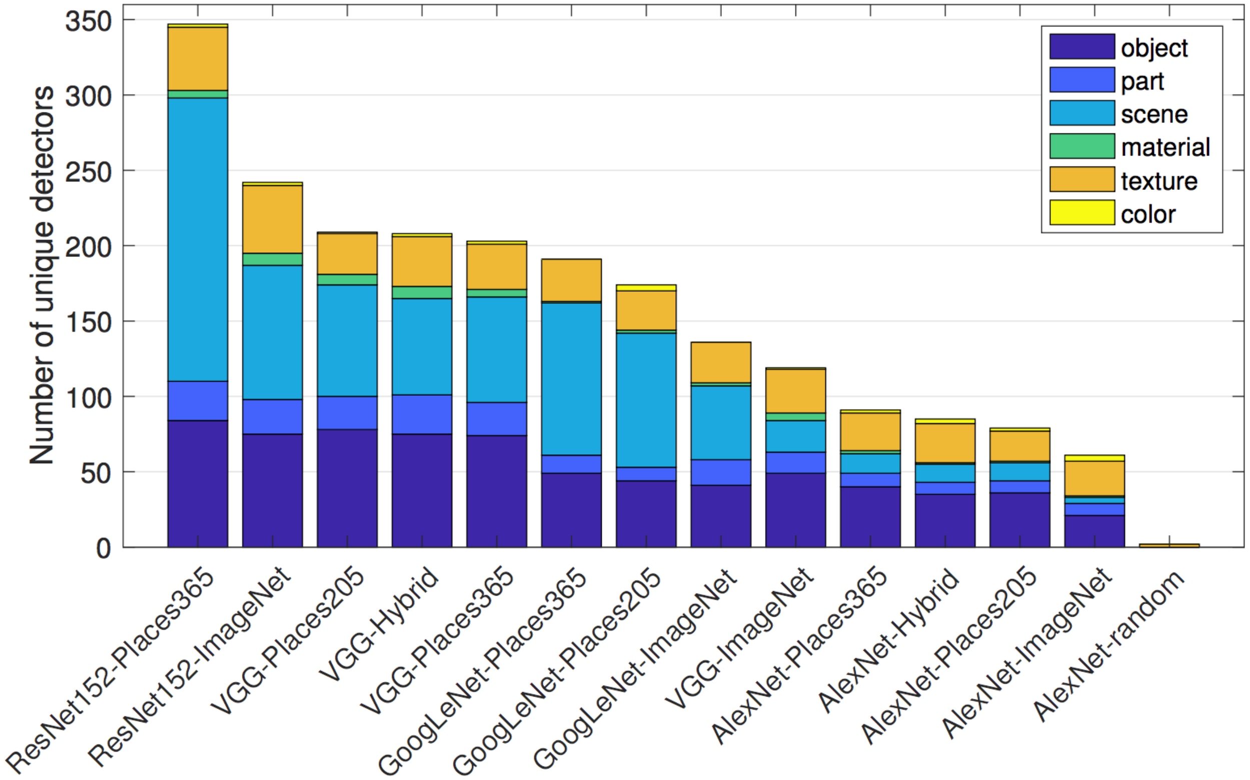
Experiments – 3. Layer levels

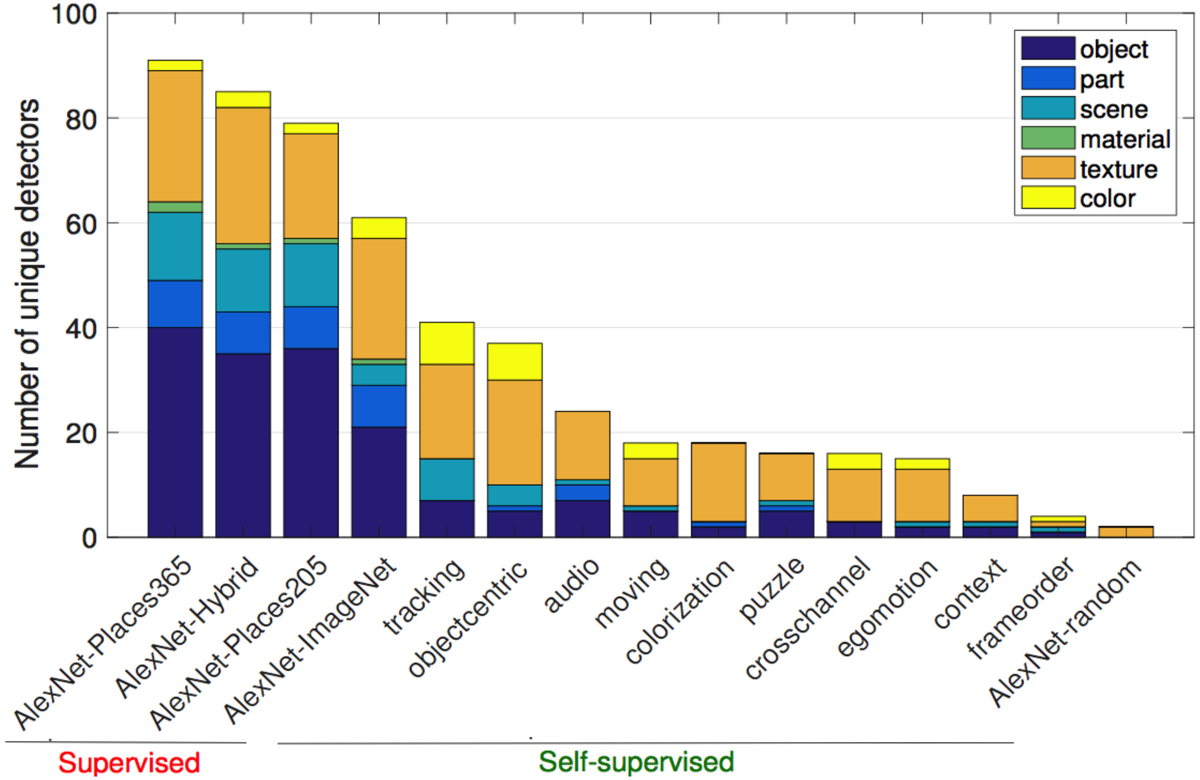




Experiments – 4. Architectures and supervisions

The unique detectors in last conv layer of each Networks

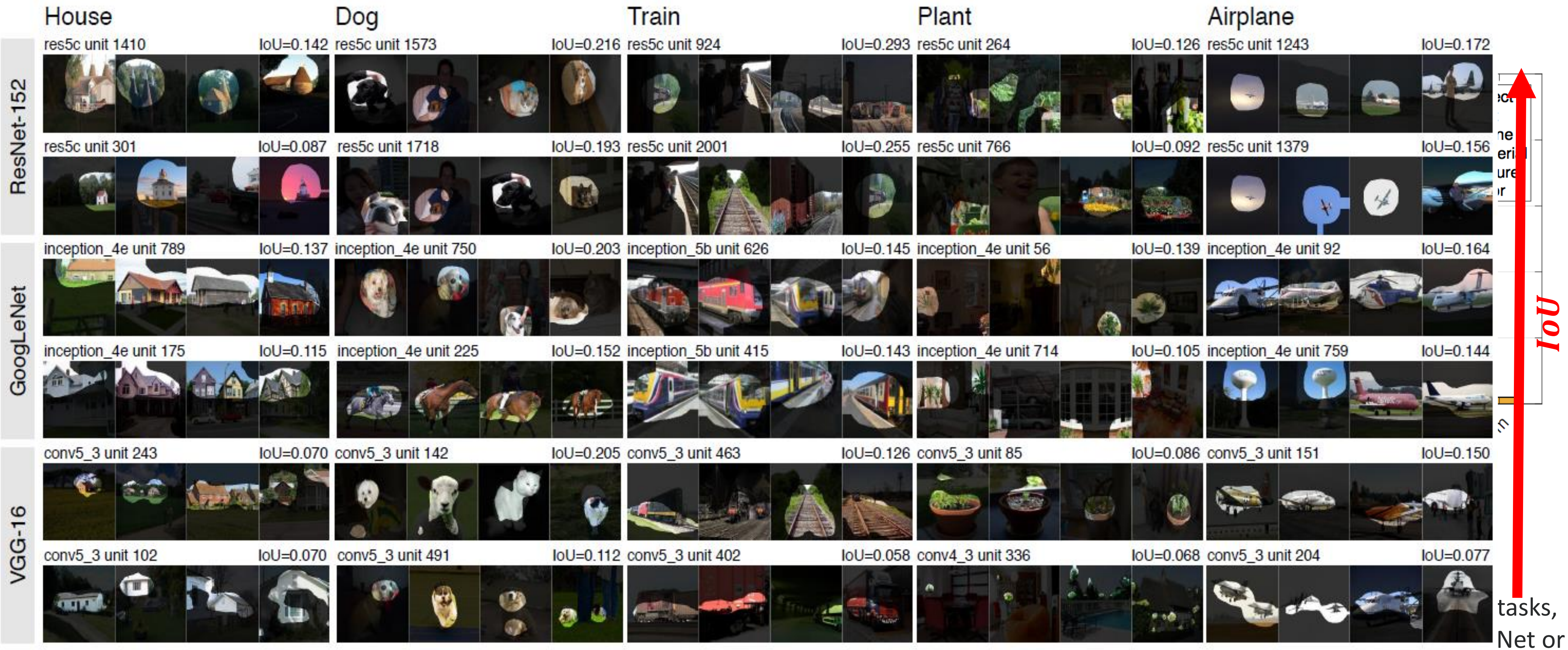




1. Interpretability of ResNet > VGPlaces205 G > GoogLeNet > AlexNet, and in terms of primary training tasks, we find Places365 > > ImageNet.

2. Interpretability varies widely under a range of self-supervised tasks, and none approaches interpretability from supervision by ImageNet or Places.

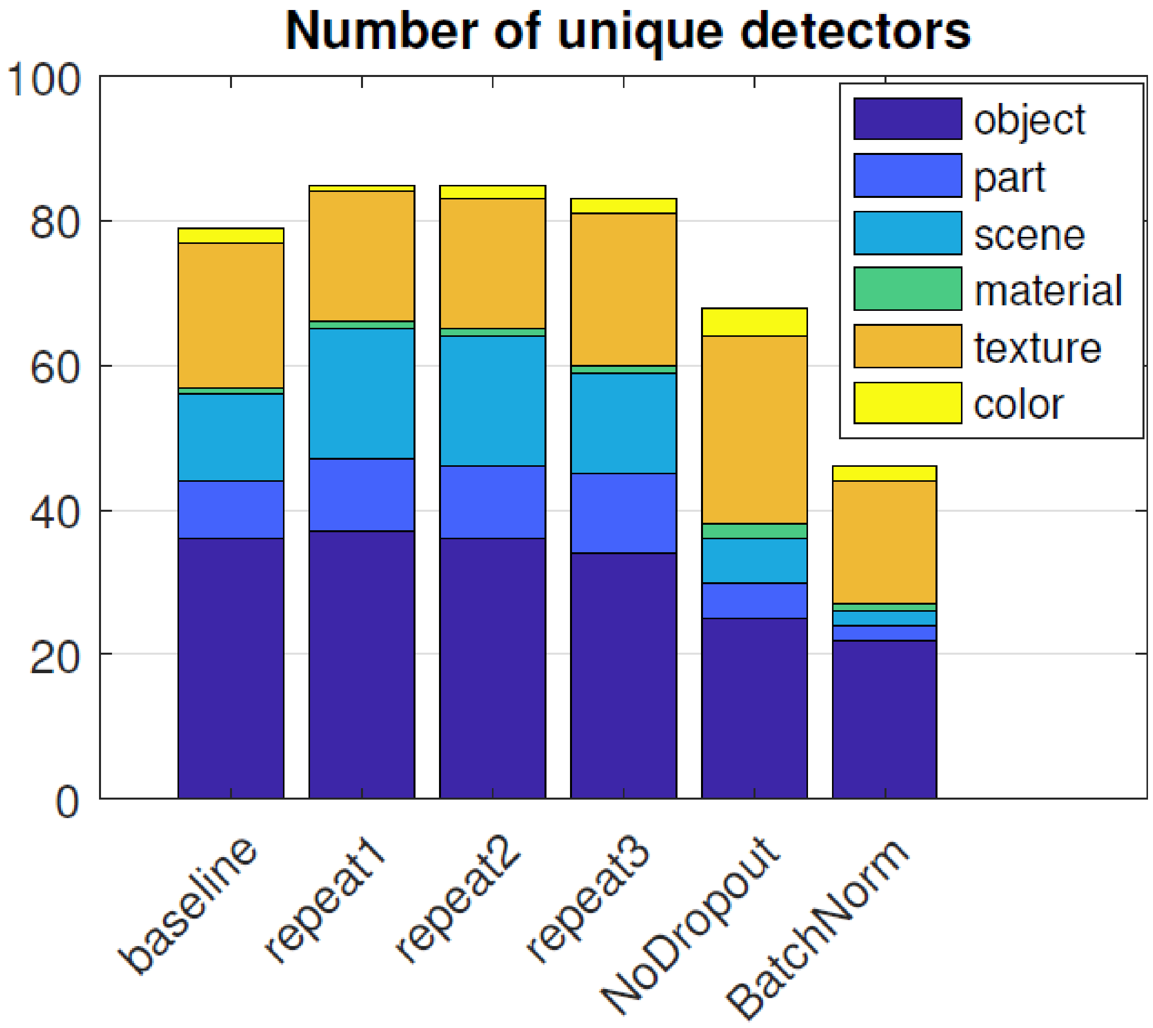
Experiments – 4. Architectures and supervisions

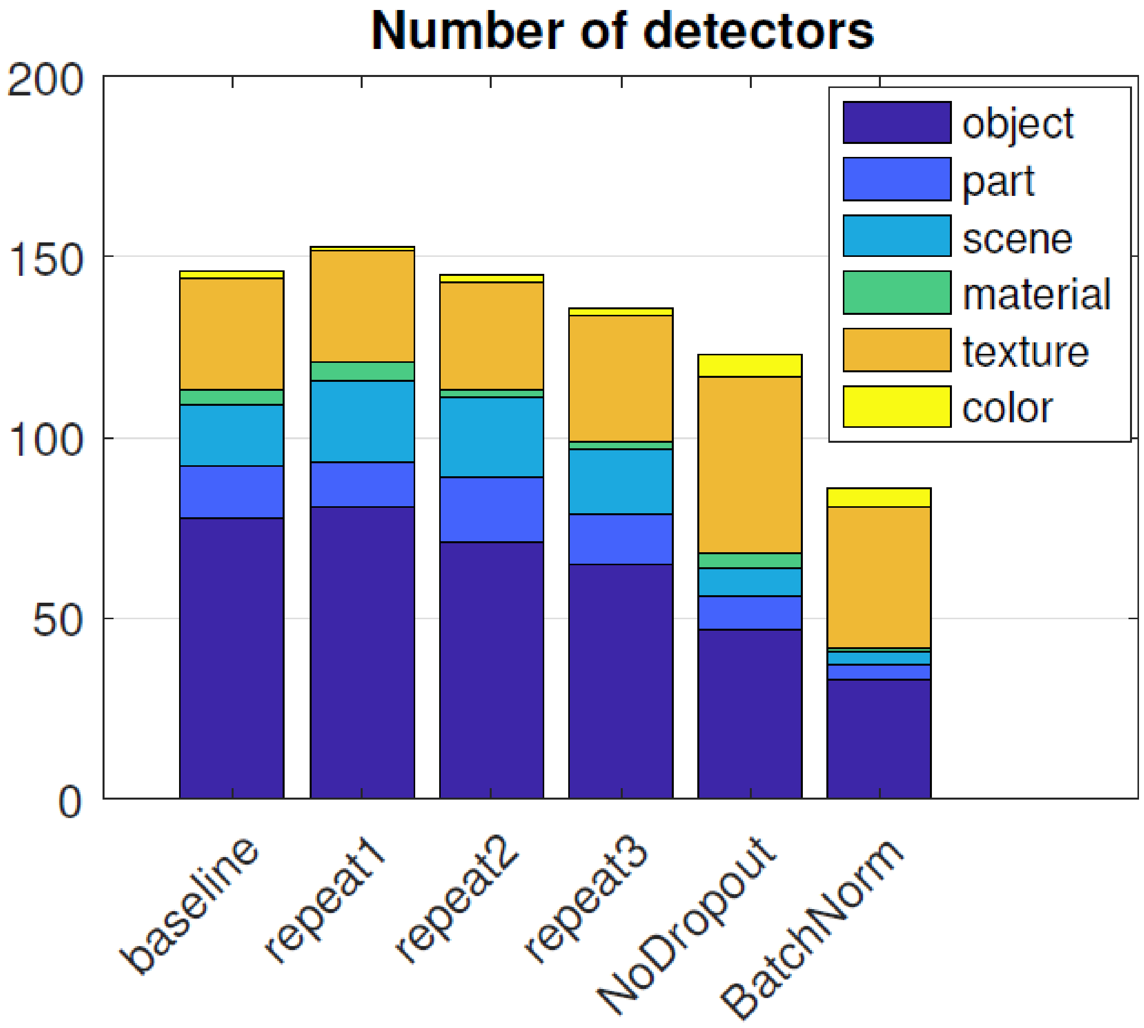


tasks,
Net or

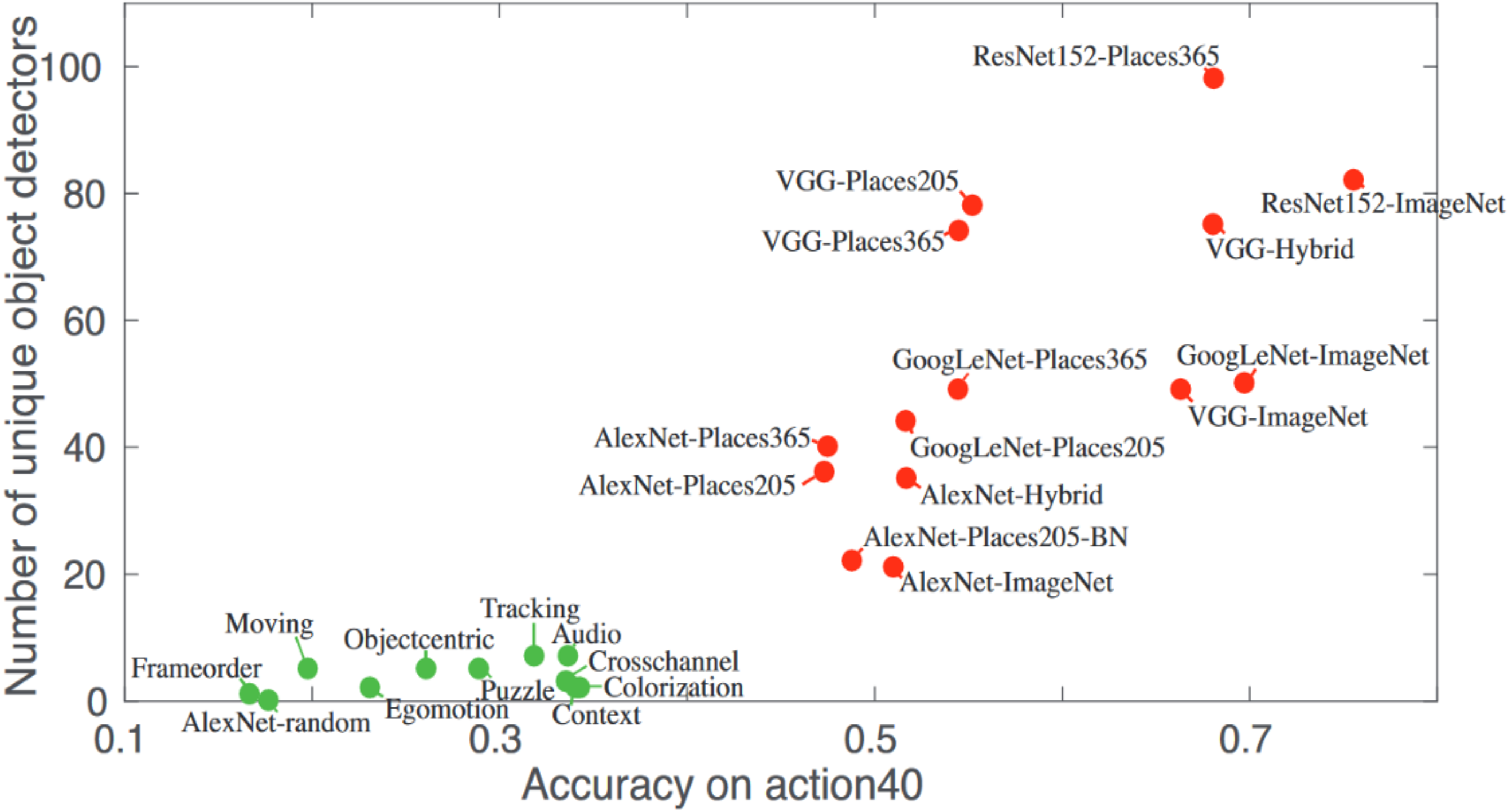
Places.

Experiments – 5. Training conditions vs. Interpretability



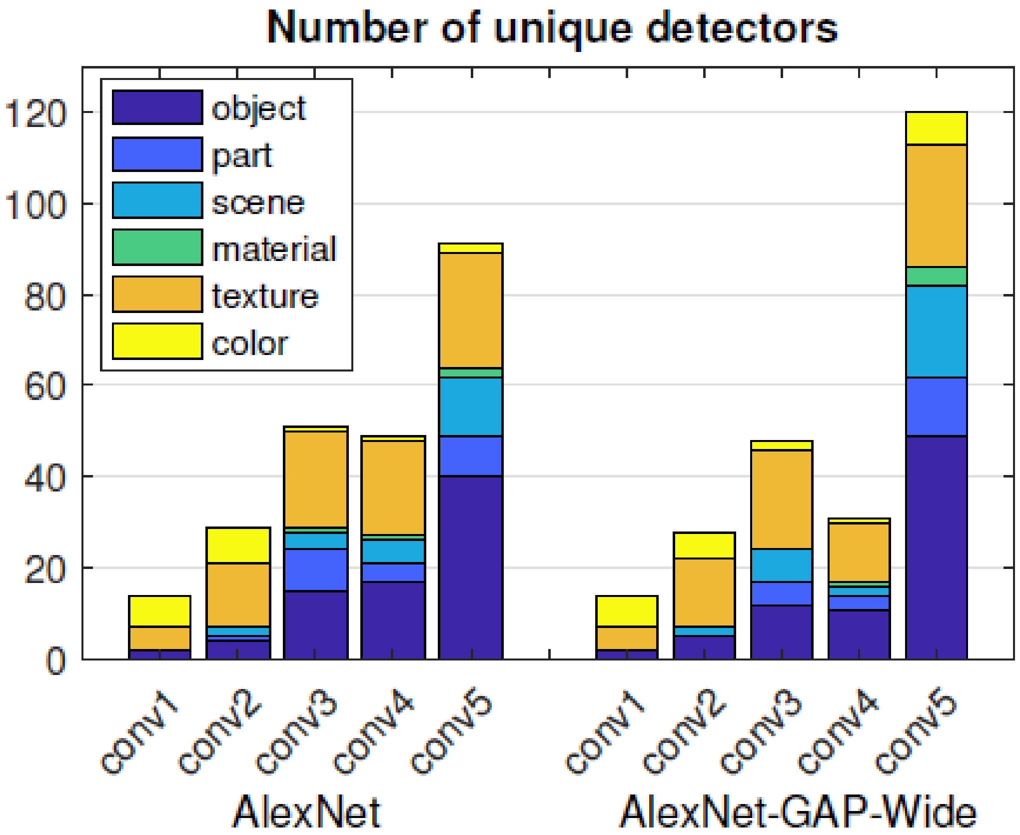


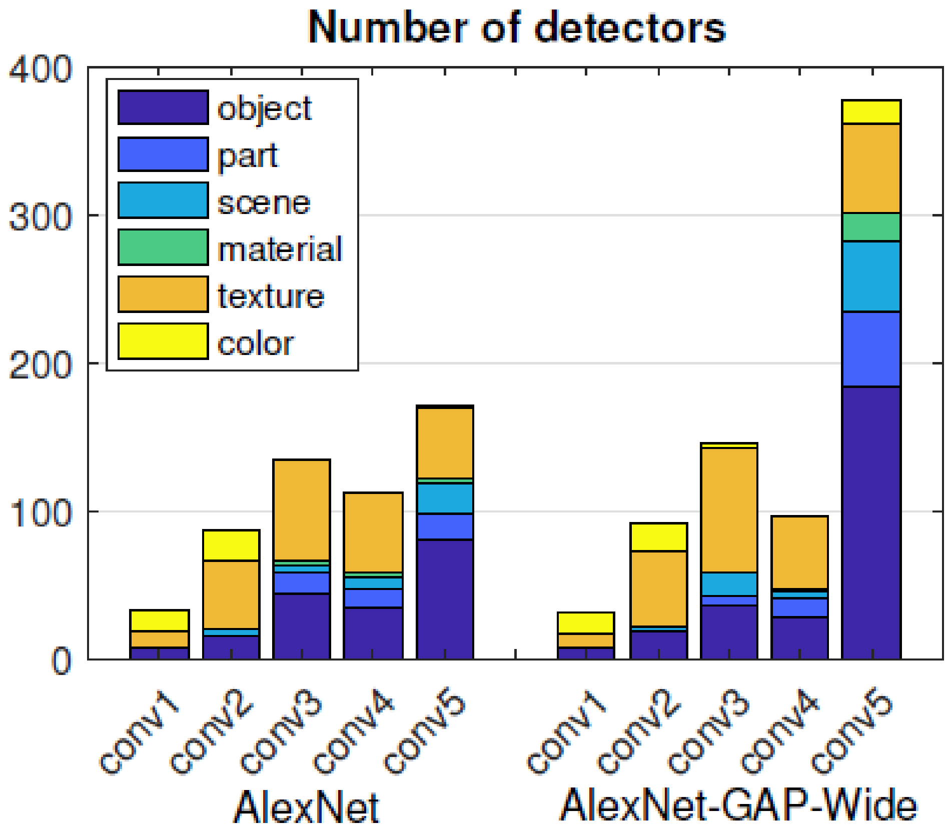
Experiments – 6. Discrimination vs. Interpretability



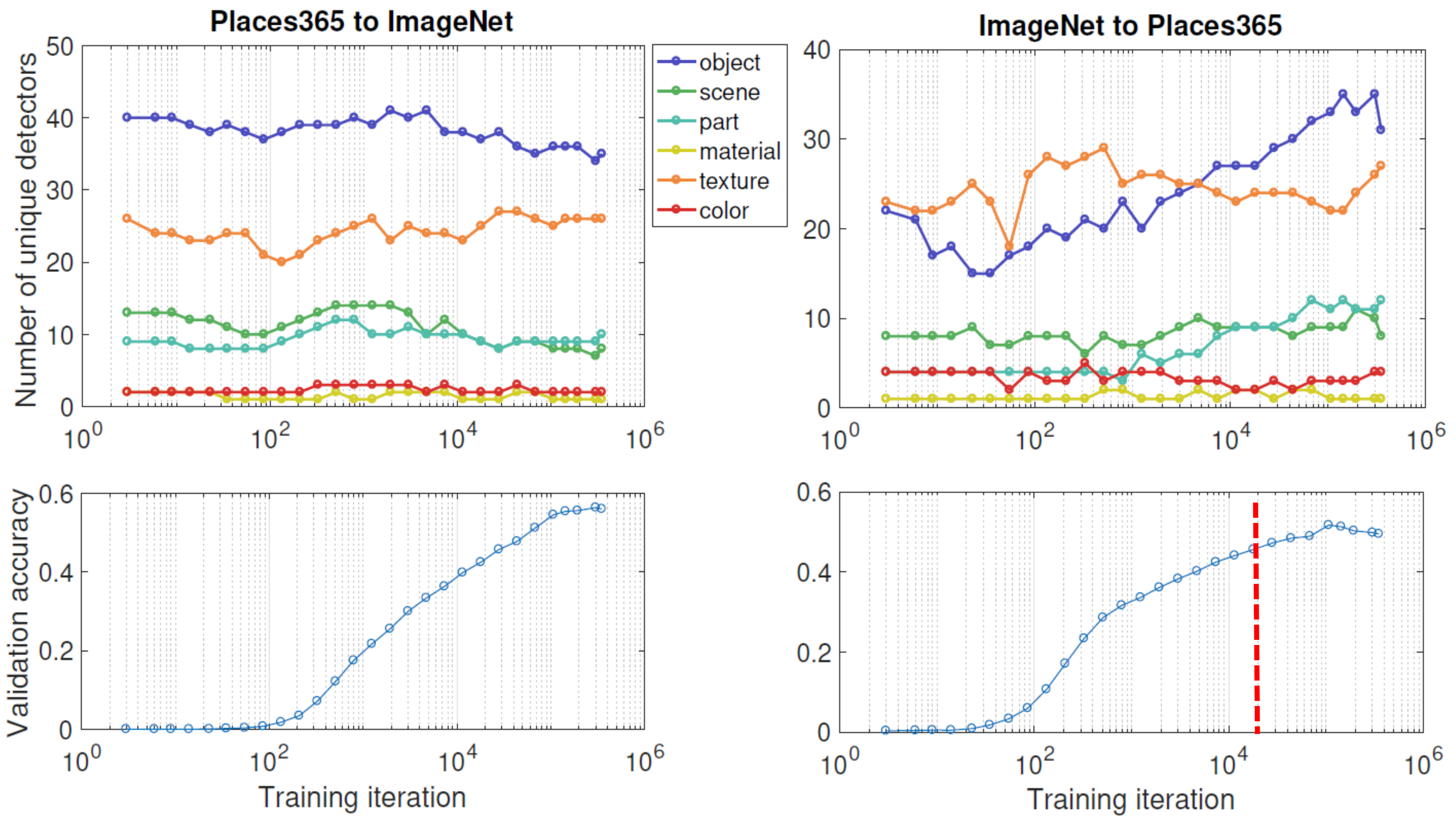
Experiments – 7. Layer Width vs. Interpretability

AlexNet-GAP-Wide: Remove FC-layers, triple the number of units in conv5, i.e. 256 to 768 units, finally put a global average pooling layer after conv5 and fully connect the pooled 768-features activations to the final class prediction.

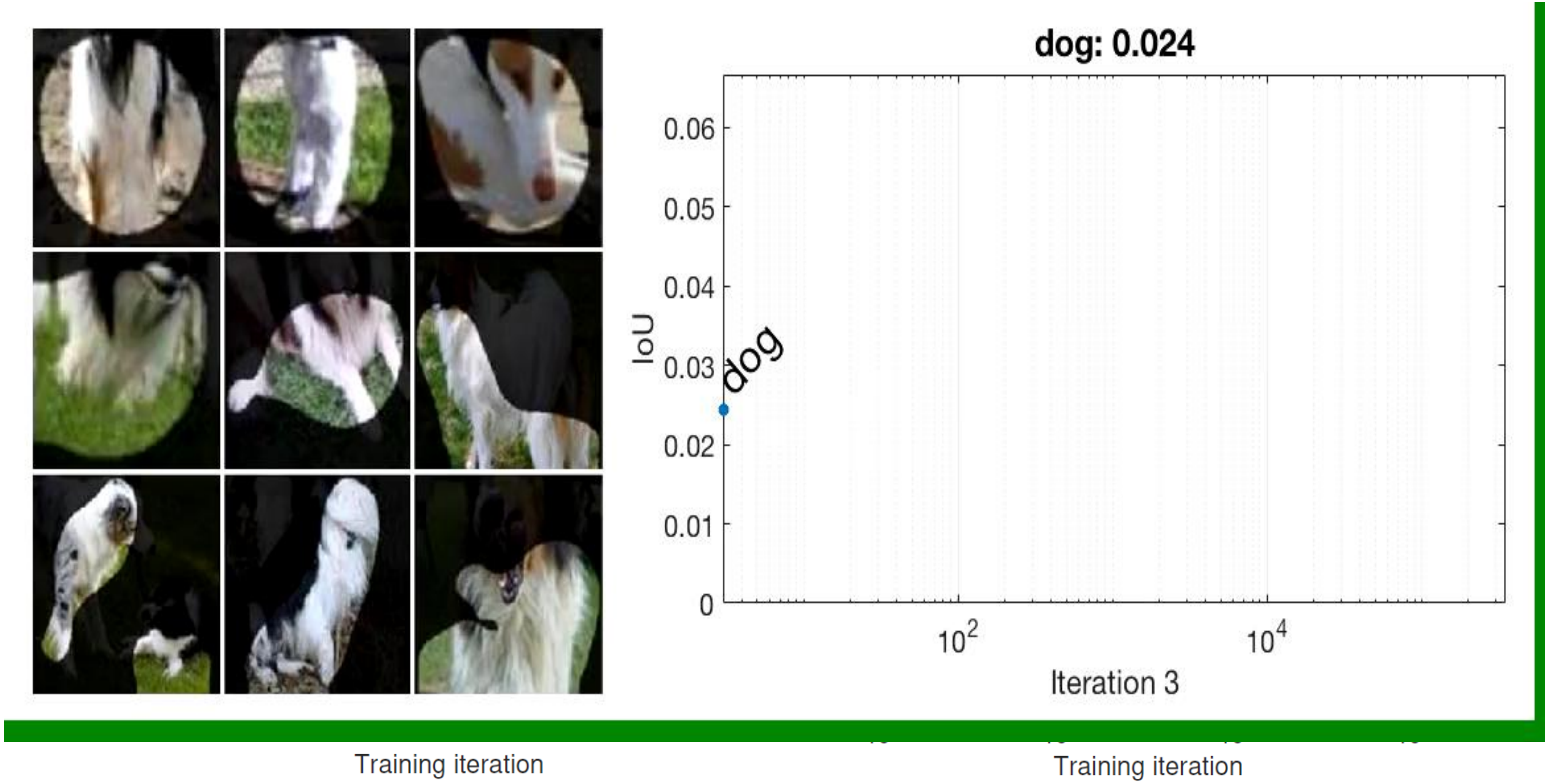




Experiments – 8. Fine-tuning



Experiments – 8. Fine-tuning



Conclusion

1. Interpretability is not an axis-indepedent phenomenon
2. Deeper CNNs architectures appear to allow a greater interpretability, which also increases with the concepts that training set contains
3. Representation at different layers of CNNs disentangle different categories of meaning
4. Different training techniques and condition lead to a significant change of interpretability of representation learned by hidden units.
5. Interpretability and discriminative power are two qualities that need to be measured separately, though they have a positive correlation.

Reference

Papers:

[1]. D. Bau, B. Zhou. 2017. Network Dissection: Quantifying Interpretability of Deep Visual Representations

[2]. B. Zhou, A. Khosla,. 2015. Object detectors emerge in deep scene cnns. International Conference on Learning Representations, 2015.

Figures & Tables:

Fig 1. https://futureoflife.org/2017/05/30/on-ai-prescription-drugs-and-managing-the-risks-of-things-we-dont-understand/

Fig 2. https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/

** All the figures and tables without number are taken from the original paper and their presentation slides, available on: http://netdissect.csail.mit.edu/*

Thank you!

Pingchuan Ma

Contact: *P.Ma@stud.uni-heidelberg.de*