Network Dissection:
Quantifying Interpretability of Deep Visual Representations

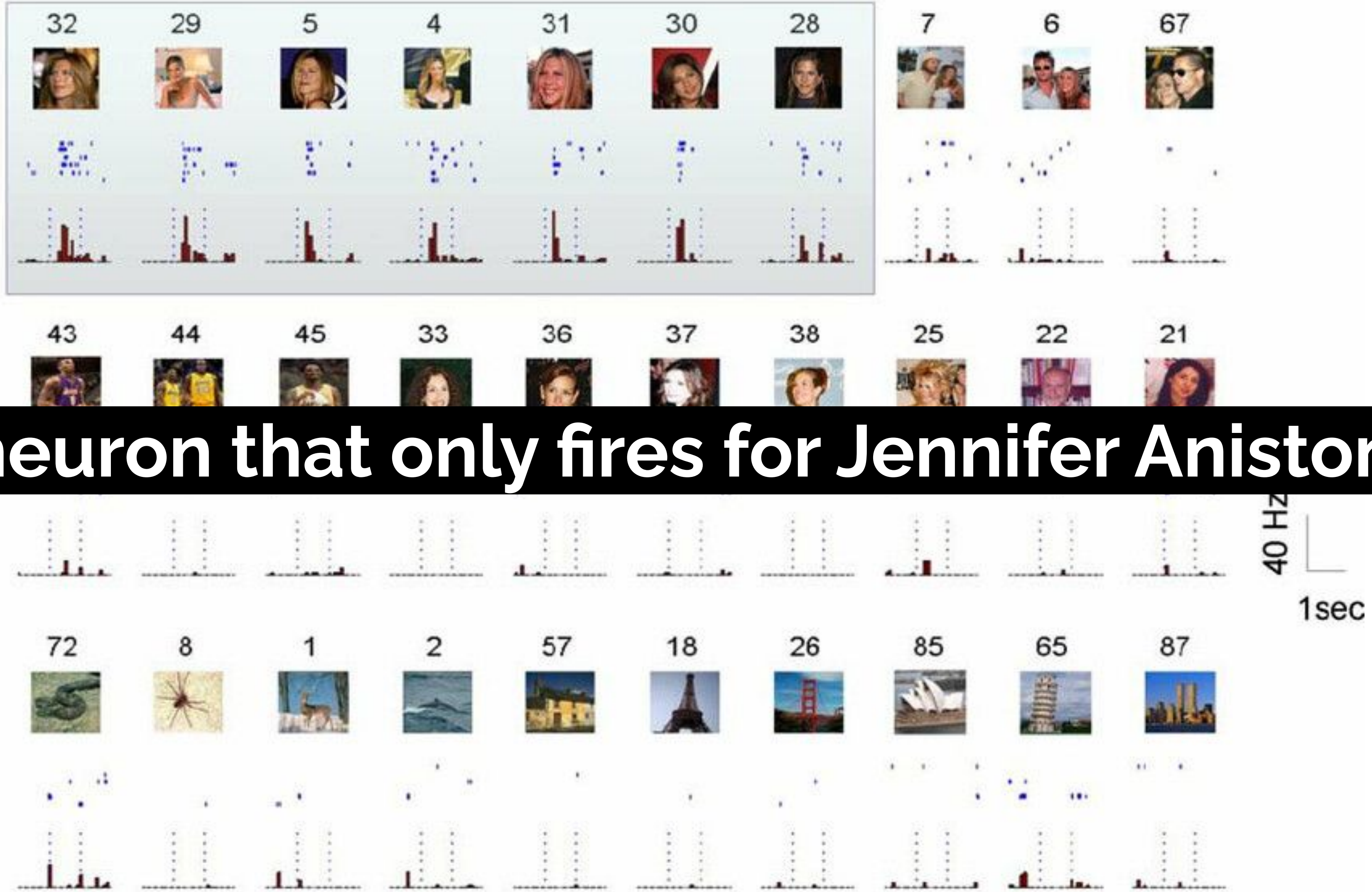
David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and, Antonio Torralba CSAIL, MIT

Presented by Anat Kleiman, Gustaf Ahdritz, Xin Tang, and Luke Bailey

Presentation Roadmap

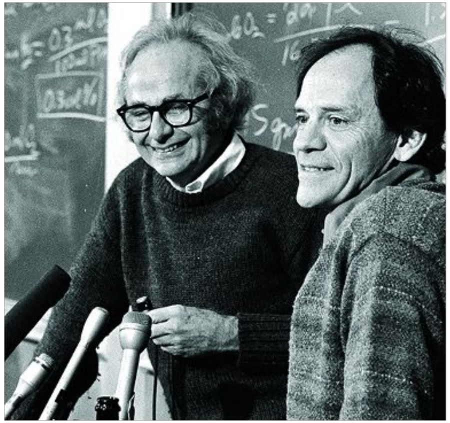
- **Introduction:**
  - Motivation
  - Questions paper aims to answer
  - Related Works
- **Method**
- **Experiments**
  - Training Conditions
  - Discrimination
  - Layer Width

How is semantic visual concept represented in the brain?

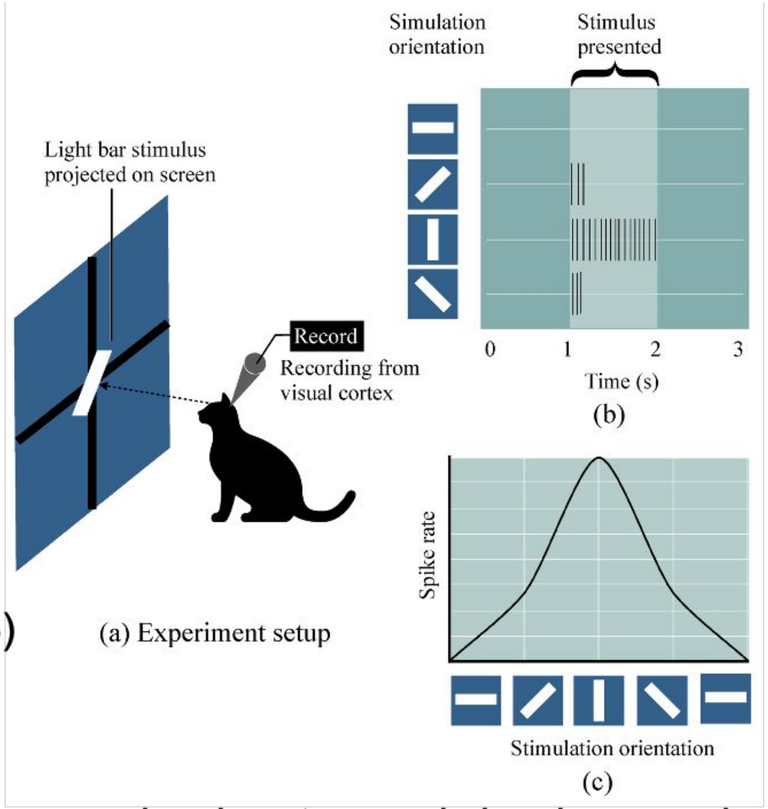


A neuron that only fires for Jennifer Aniston

Disentangled representation in visual cortex

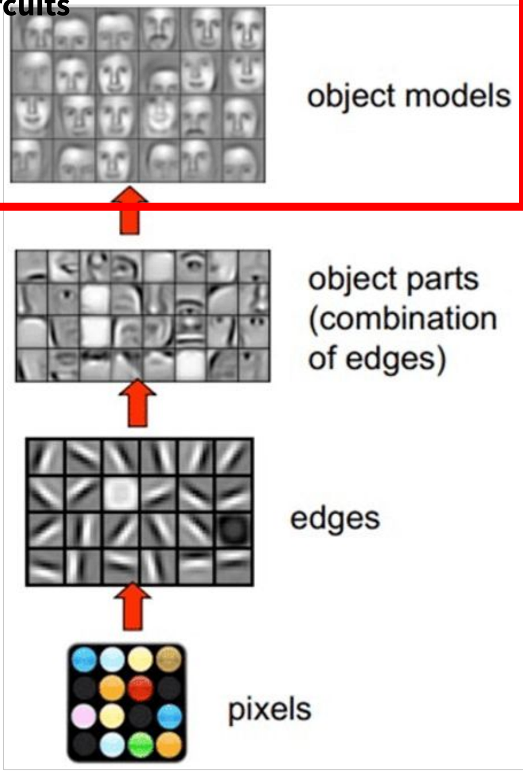


David Hubel (1926-2013)
Torsten Wiesel (1924-)



visual system and visual processing
Nobel Prize 1981

Jennifer Aniston neuron is in high-level neural circuits



Hierarchical Coding

Deep CNN for computer vision

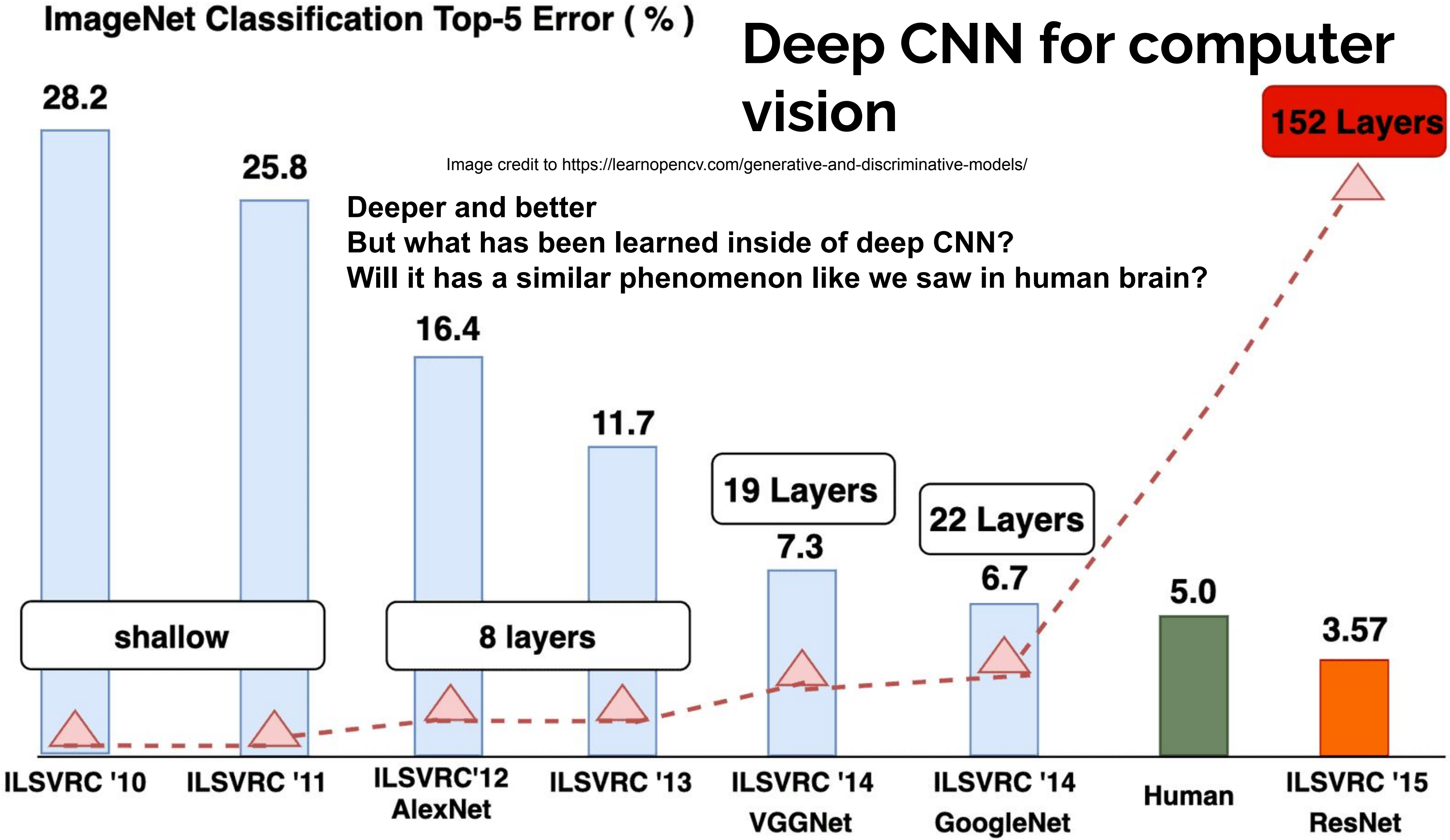


Image credit to https://learnopencv.com/generative-and-discriminative-models/

Deeper and better
But what has been learned inside of deep CNN?
Will it has a similar phenomenon like we saw in human brain?

Proposed questions

1. What is a disentangled representation, and how can its factors be quantified and detected (in deep CNN)?

2. Do interpretable hidden units reflect a special alignment of feature space, or are interpretations a chimera?

3. What conditions in state-of-the-art training lead to representations with greater or lesser entanglement?

Proposed questions and contributions

1. **What is a disentangled representation, and how can its factors be quantified and detected?**
    a. Proposed a metric, intersection over union score (IoU), to quantify the interpretability of each unit
    b. The alignment level between unit activated area and human-interpretable concepts
2. **Do interpretable hidden units reflect a special alignment of feature space, or are interpretations a chimera?**
    a. A semantic concept can be detected by many units
    b. A unit can detect many semantic concepts
3. **What conditions in state-of-the-art training lead to representations with greater or lesser entanglement?**
    a. Number of unique detectors, Layer depth, training iterations
    b. the angle of the images, input datasets
    c. Fine-tuning, supervised v.s. unsupervised

Related works

1. **Generative Visualizations of Individual Units**
    - Mahendran et al., CVPR 2015
    - Nguyen et al., NIPS 2016
    - Simonyan et al., ICML 2014
2. **Salience-based Visualizations of Individual Units**
    - Deconvolution: Zeiler et al., ECCV 2014
3. **Visualizing Representations as a Whole**
    - t-SNE: Maaten et al., JMLR, 2008
    - prototype autoencoder: Li et al., AAAI, 2018
    - Yosinski et al., ICML, 2015 …

Limitation of these works: qualitative analyses, cannot be used for comparison between models

Presentation Roadmap

- Introduction
- **Method**
- Experiments

Broden: Broadly and Densely Labeled Dataset

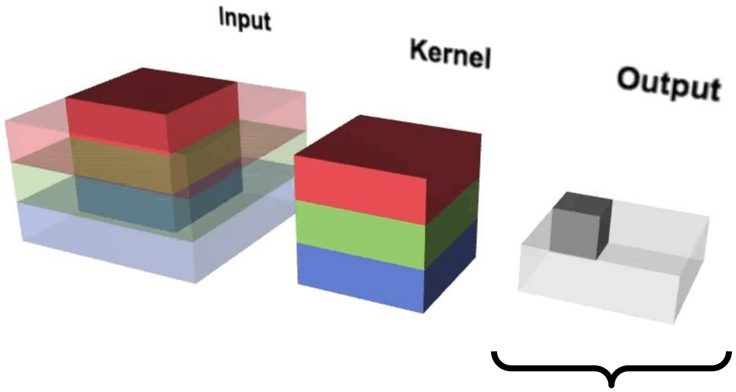
- Combination of multiple datasets with segmentation and image wide labels

Figure 2. Samples from the **Broden** Dataset. The ground truth for each concept is a pixel-wise dense annotation.

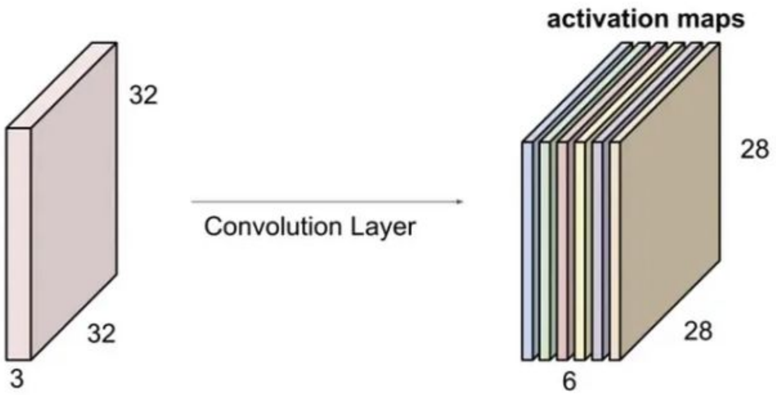
- Multiple labels can apply to the same pixel e.g. “cat, leg, black”.

Scoring Unit Interpretability

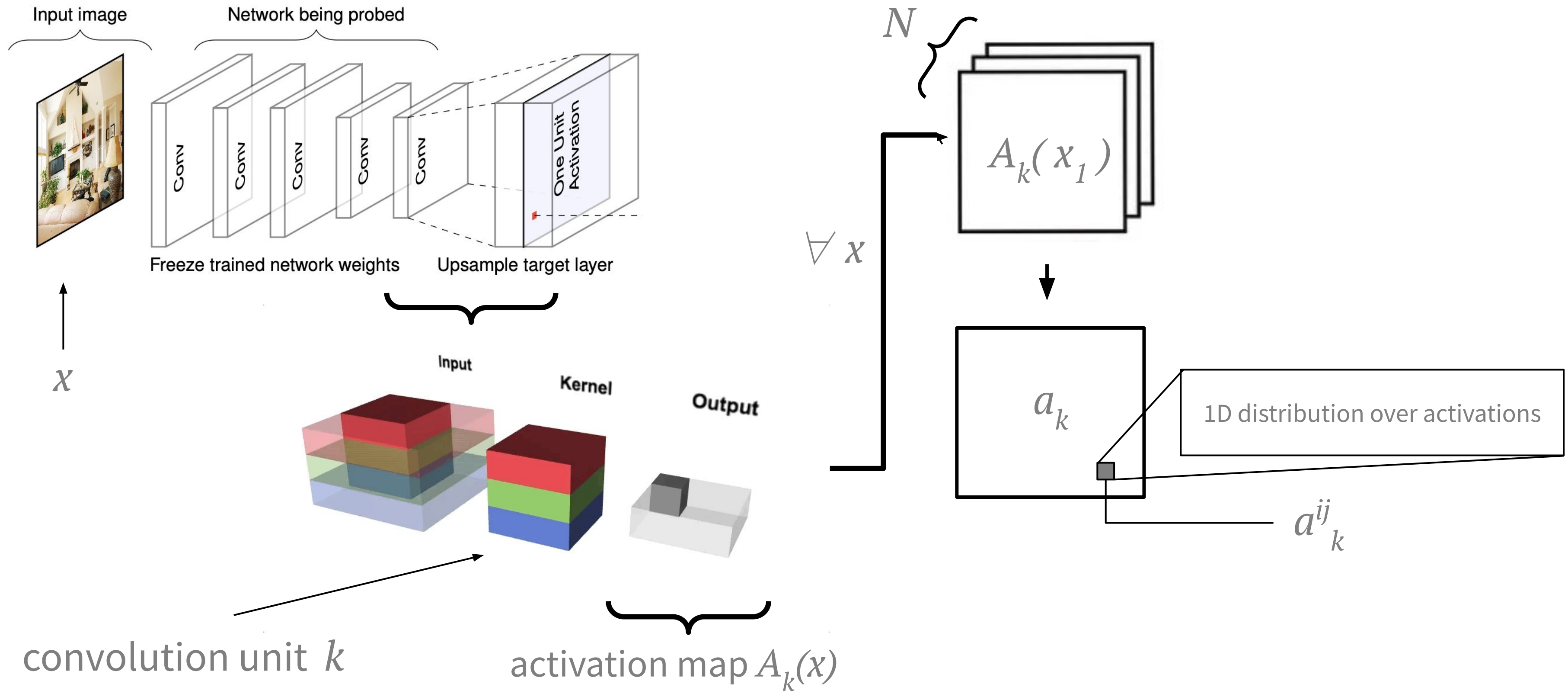
- Unit is a convolutional filter



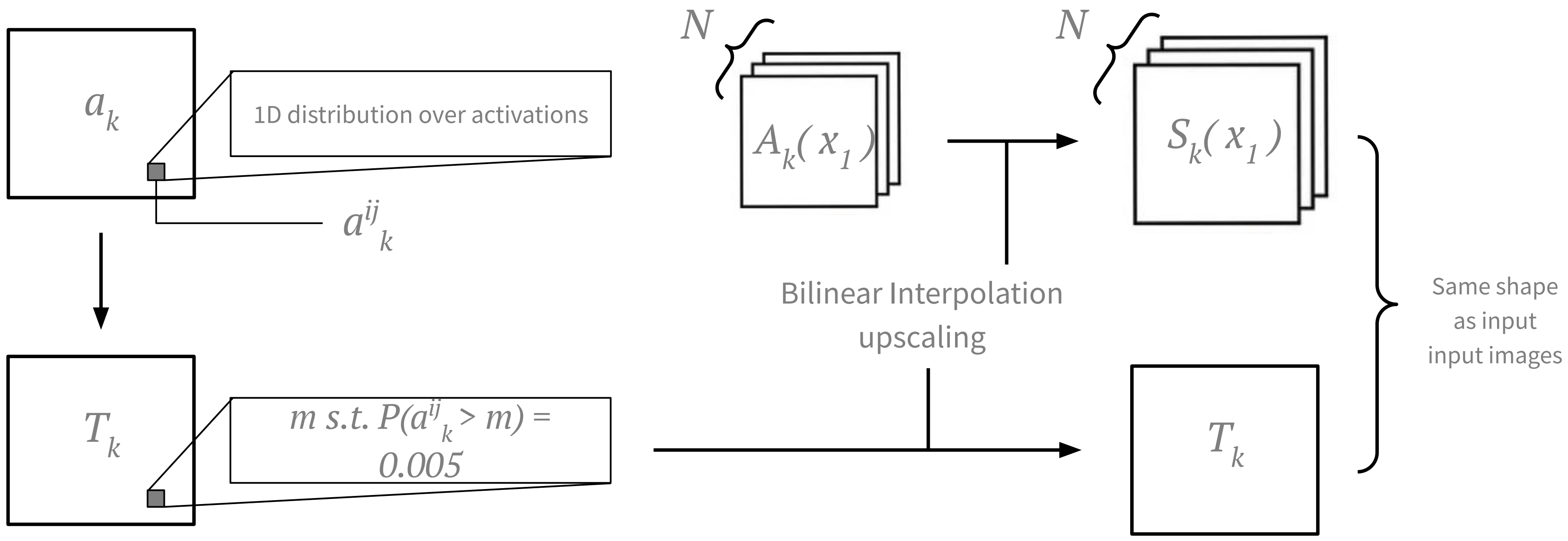
activation map



Scoring Unit Interpretability



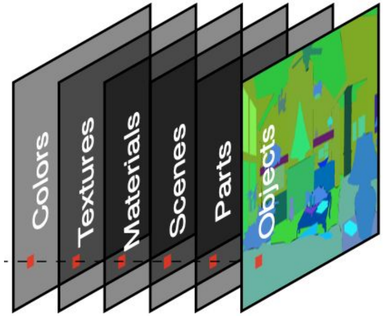
Scoring Unit Interpretability



Scoring Unit Interpretability

$$\overbrace{S_k(x_1)}^{N} \geq T_k = \overbrace{M_k(x_1)}^{N}$$

L_c binary concept maps



$$IoU_{k,c} = \frac{\sum |M_k(\mathbf{x}) \cap L_c(\mathbf{x})|}{\sum |M_k(\mathbf{x}) \cup L_c(\mathbf{x})|}$$

The value of $IoU_{k,c}$ is the accuracy of unit k in detecting concept c. Consider one unit k as a detector for concept c if $IoU_{k,c}$ exceeds a threshold. They use $IoU_{k,c}$ > 0.04

Scoring Unit Interpretability

$$IoU_{k,c} = \frac{\sum |M_k(\mathbf{x}) \cap L_c(\mathbf{x})|}{\sum |M_k(\mathbf{x}) \cup L_c(\mathbf{x})|}$$

- Changing *IoU* threshold changes number of concept detectors but not orderings between networks.
- One unit might be detector for multiple concepts, they just choose the top ranked concept for an individual unit.
- Interpretability of a layer is the number of unique concepts aligned with units

Presentation Roadmap

- Introduction
- Method
- Experiments

Experiments

Scene-centric data set with categories
such as kitchen, living room, and coast

Table 2. Tested CNNs Models

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.

Experiment 1: Human Evaluation of Interpretations

1. **Identify Interpretable units** - units that raters agreed with ground-truth interpretations from [*]
2. Raters shown 15 images with highlighted patches showing the most highly-activating regions for each unit in AlexNet trained on Places205, and asked to decide (yes/no) whether a given phrase describes most of the image patches.
3. **Find network dissection** - the portion of interpretations generated by method that were rated as descriptive
4. **Human Consistency** - portion of ground-truth labels that were found to be descriptive by a second group of raters

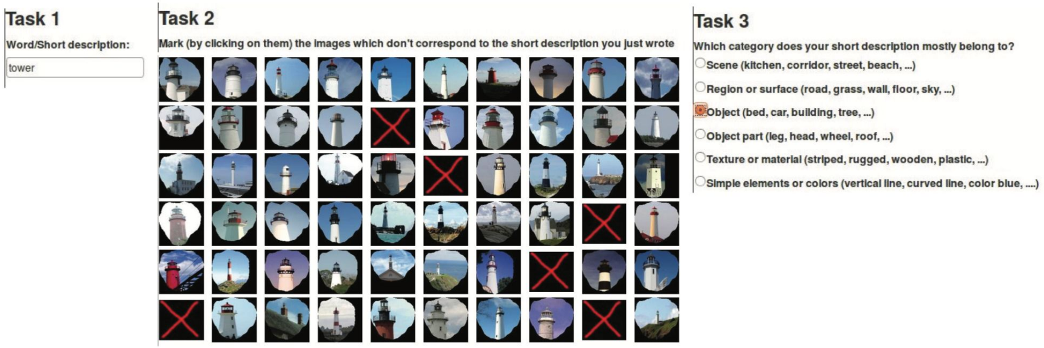


Figure 6: AMT interface for unit concept annotation. There are three tasks in each annotation.

[*] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Object detectors emerge in deep scene cnns. International Conference on Learning Representations, 2015.

Experiment 1: Human Evaluation of Interpretations

Table 3. Human evaluation of our Network Dissection approach. Interpretable units are those where raters agreed with ground-truth interpretations. Within this set we report the portion of interpretations assigned by our method that were rated as descriptive. Human consistency is based on a second evaluation of ground-truth labels.

	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%

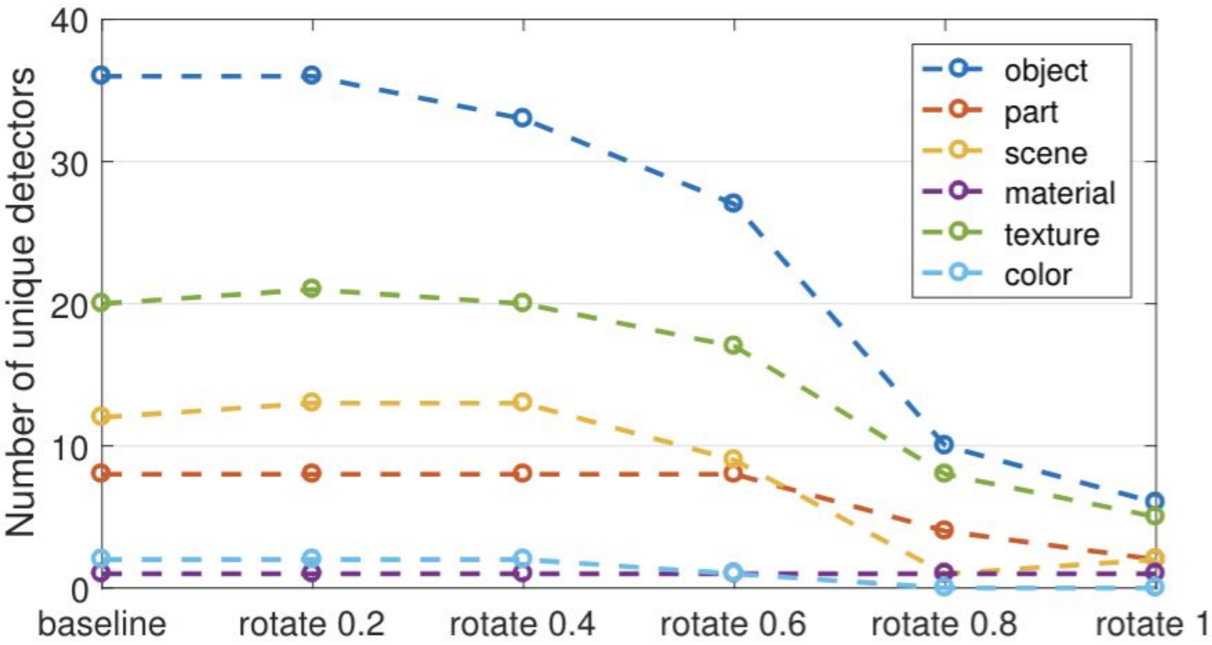
Experiment 2: Axis-Aligned Interpretability

Two hypotheses:

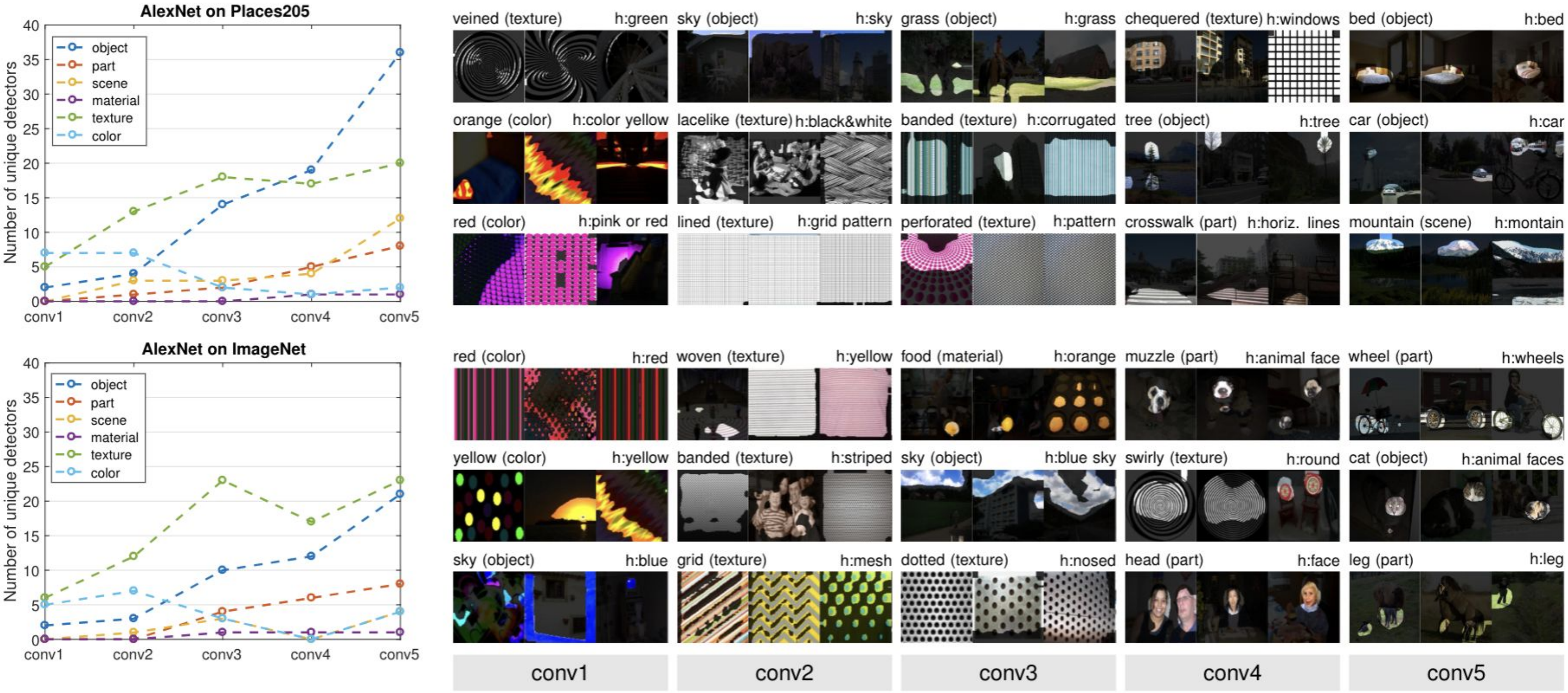
1) **Concepts appear in every direction**
    - **Default hypothesis.** Single units not much more interpretable than combinations of units.

2) **Concepts are rare + the model converges to a special, semantically rich basis**
    - The model’s natural basis is a meaningful decomposition.

Experiment 2: Axis-Aligned Interpretability



Experiment 3: Concepts by layer



Experiment 4: Network Architectures

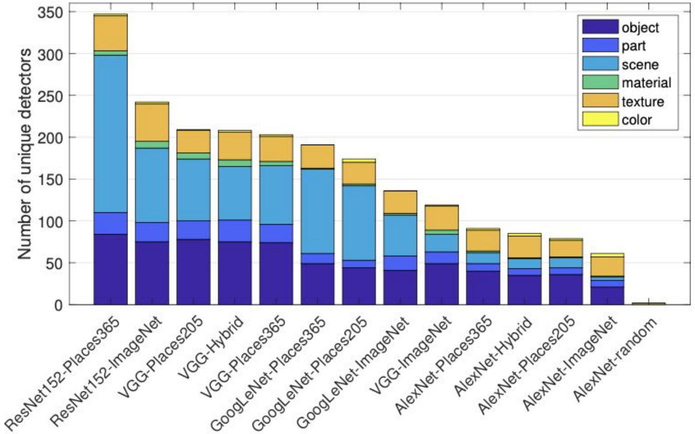


Figure 7. Interpretability across different architectures and training.

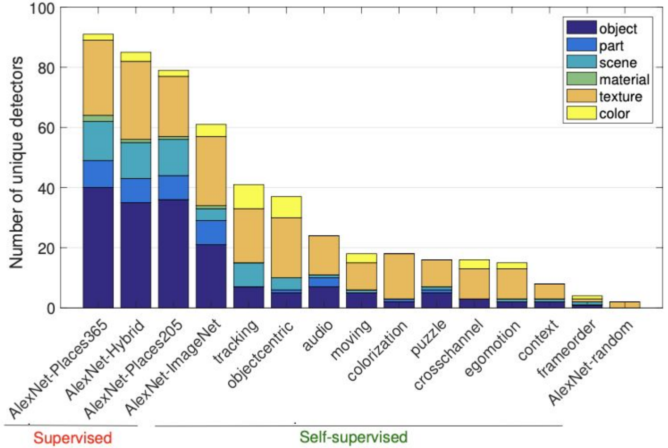


Figure 8. Semantic detectors emerge across different supervision of the primary training task. All these models use the AlexNet architecture and are tested at conv5.

Experiment 5: Training Conditions

Varied training conditions:

1) Weight initializations

- **Minimal Effect:** Models converge to similar levels of interpretability

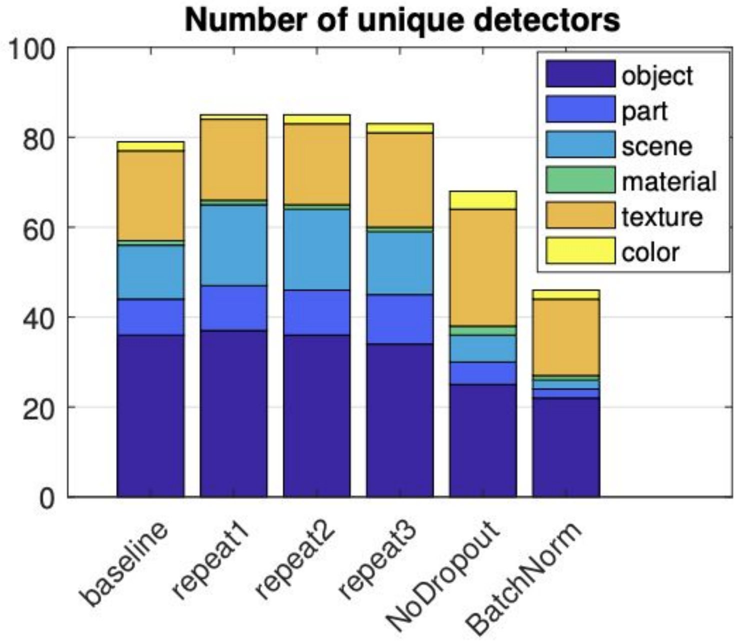
2) Dropout

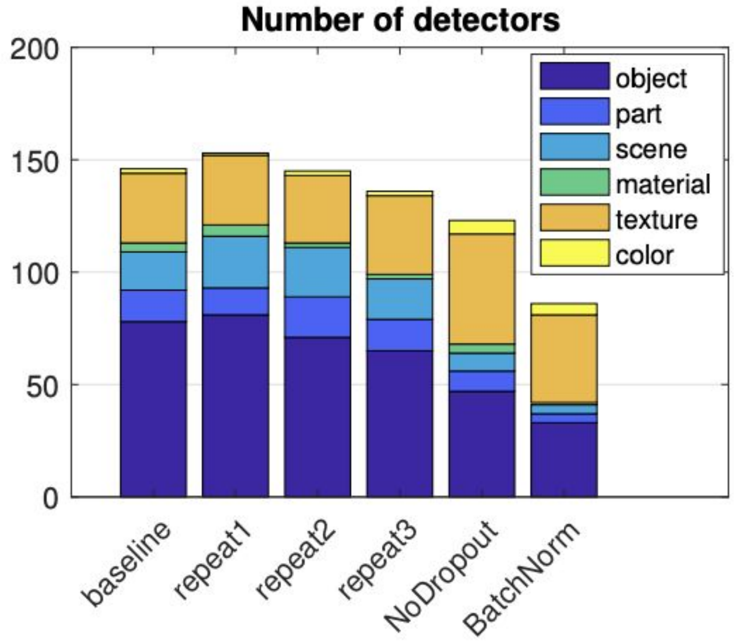
- **Some Effect:** Lack of dropout leads to more “texture” and less “object” detectors,

3) Batch Normalization

- **Significant Effect:** Interpretability decreased significantly

Experiment 5: Training Conditions



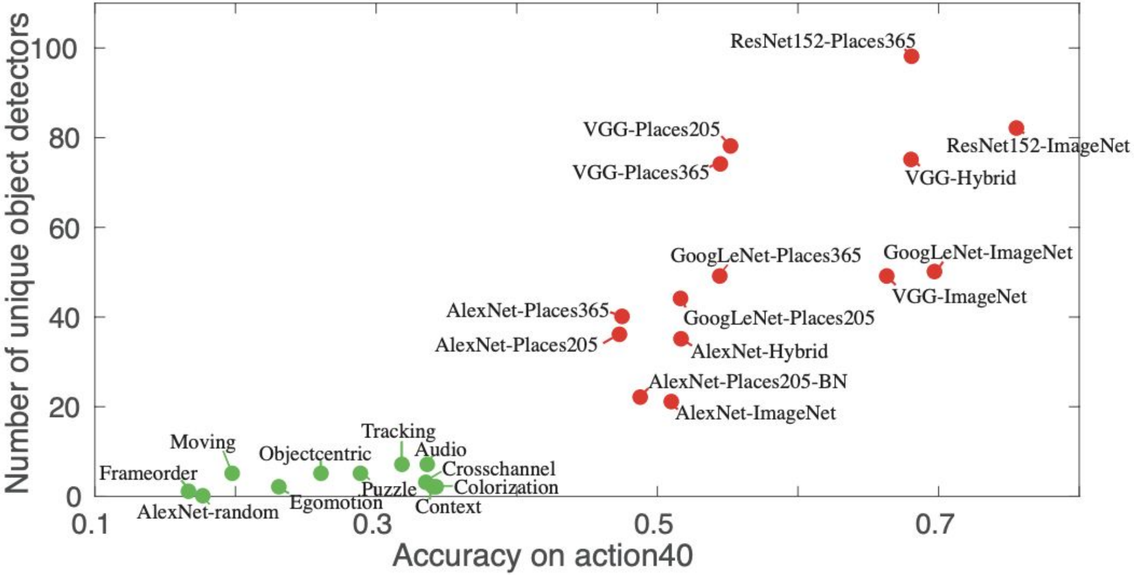


Experiment 6: Discrimination

Benchmark high-level activations on a new task:

- Across several Deep NNs, extract activation from high CNN layers
- Train a linear SVM on a new *action recognition task*
- Compute classification accuracy

Experiment 6 Discrimination



Result: Positive correlation between object detectors and classification accuracy → Encouraging **concept detection** can improve **discrimination**

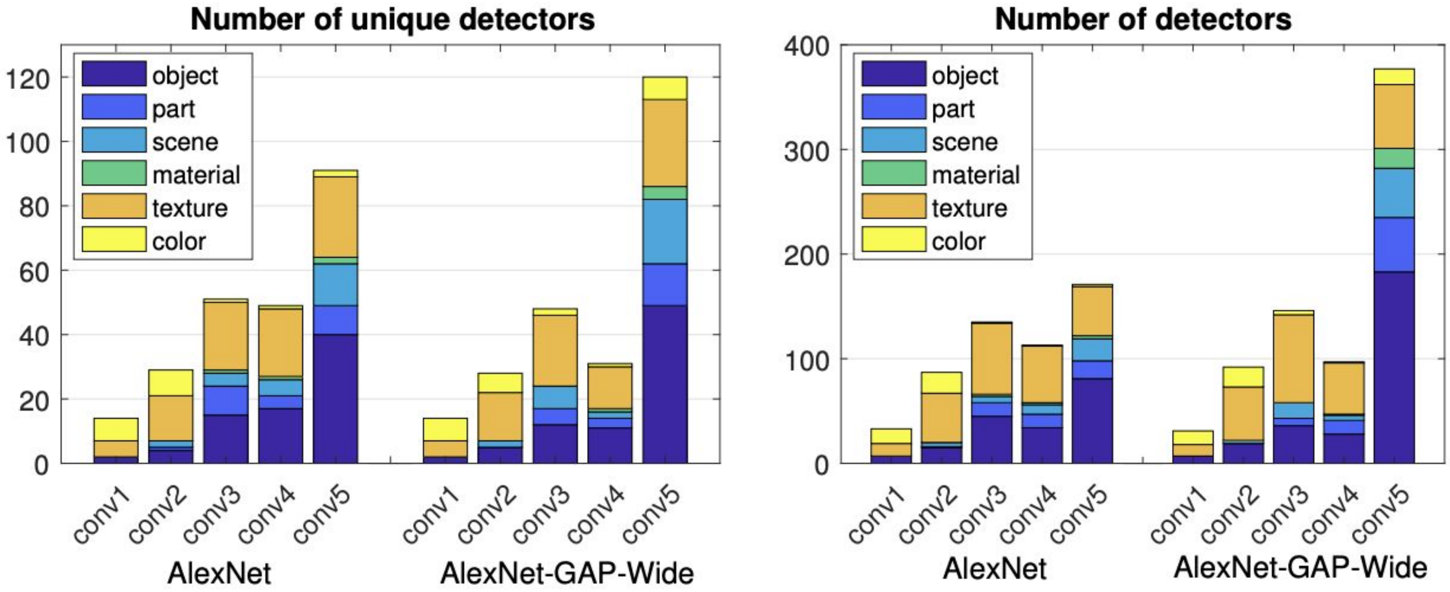
Experiment 7: Width

Effect of layer width (number of units in a layer):

Increased layer width retains similar accuracy, but much more **concept detectors**

- # Detectors increased both at increased layer and in network generally
- Increase has a threshold

Experiment 7: Width



Discussion Questions

1) **Distribution Understanding:** Concept detectors from a particular dataset betray something about the underlying distribution. How do you think this can be applied in the real world (e.g. bias detection)?

2) **Single unit to circuit:** This method interprets single units, could it be extended to larger circuits and would this be useful?

3) **Beyond vision:** This method is deeply tied to the vision domain. Could it be extended to other domains such as natural language?