Network Dissection:
Quantifying Interpretability of Deep Visual Representations

David Bau,* Bolei Zhou,* Aditya Khosla, Aude Oliva, and Antonio Torralba
CSAIL, MIT
{davidbau, bzhou, khosla, oliva, torralba}@csail.mit.edu

Slides by Nolan Dey

Motivation

- Neural networks are often treated as a black box

- Network dissection attempts to describe what features individual neurons are focusing on

Network Dissection

1. Identify a broad set of human-labeled visual concepts

2. Gather hidden variables' response to known concepts

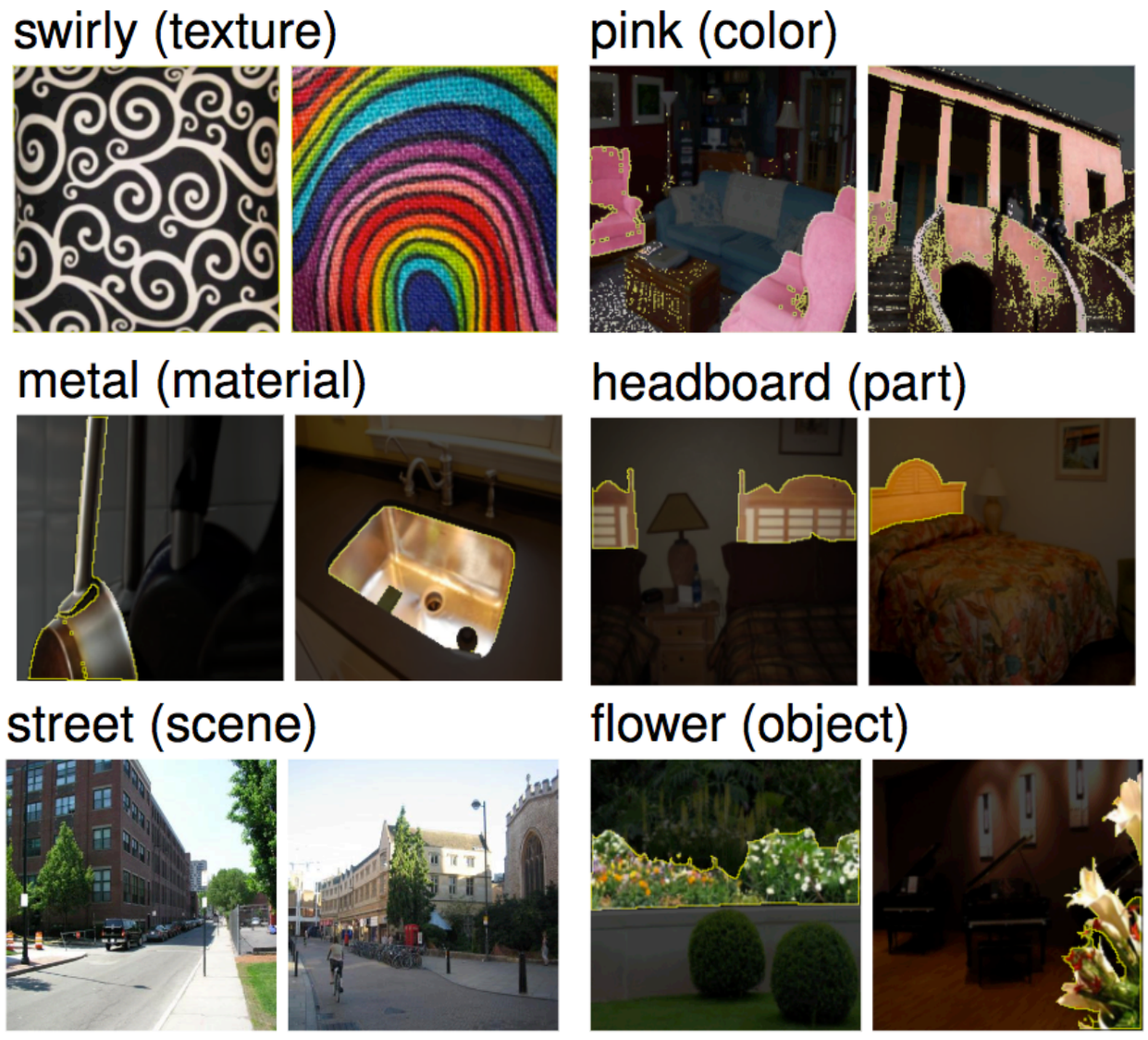
3. Quantify alignment of hidden variable - concept pairs

1. Identify a broad set of human-labeled visual concepts

- Broden dataset: Broadly and densely labelled dataset

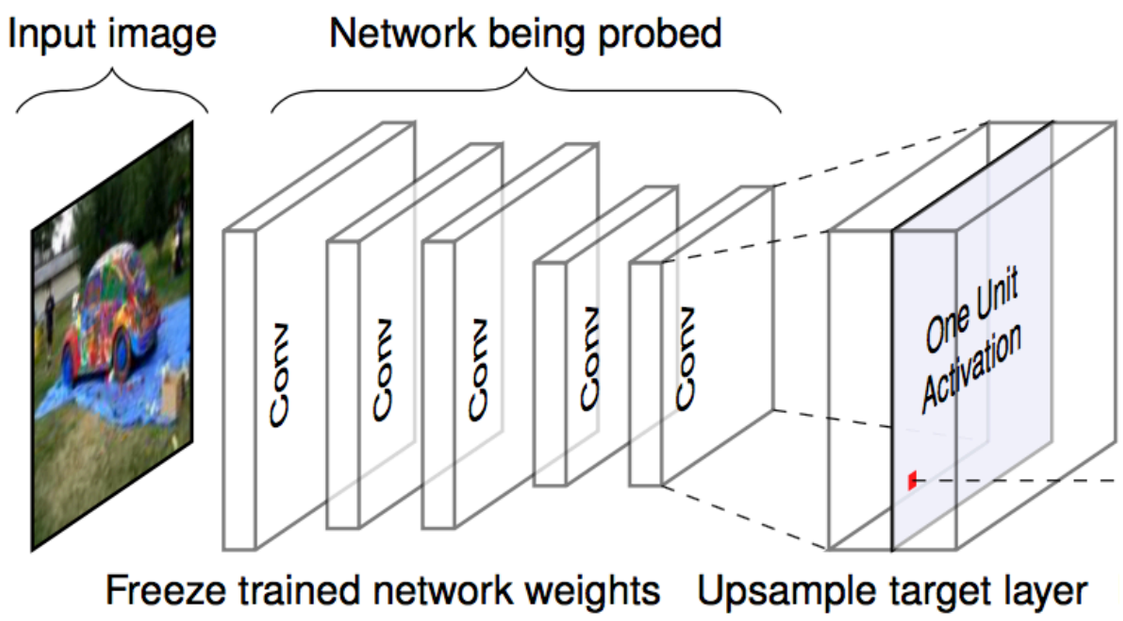
- 63,305 images with 1197 visual concepts

- Concept labels are assigned pixel-wise



2. Gather hidden variables' response to known concepts

- For convolutional neurons, compute their activation map
- In other words, what is the output of a particular convolutional filter for a given image
- Threshold this activation map to convert it to a binary activation map



3. Quantify alignment of hidden variable - concept pairs

- Measure the IoU between the binary activation map and the labelled concept images

- If activation map overlaps highly with a concept, the neuron is a detector for that concept

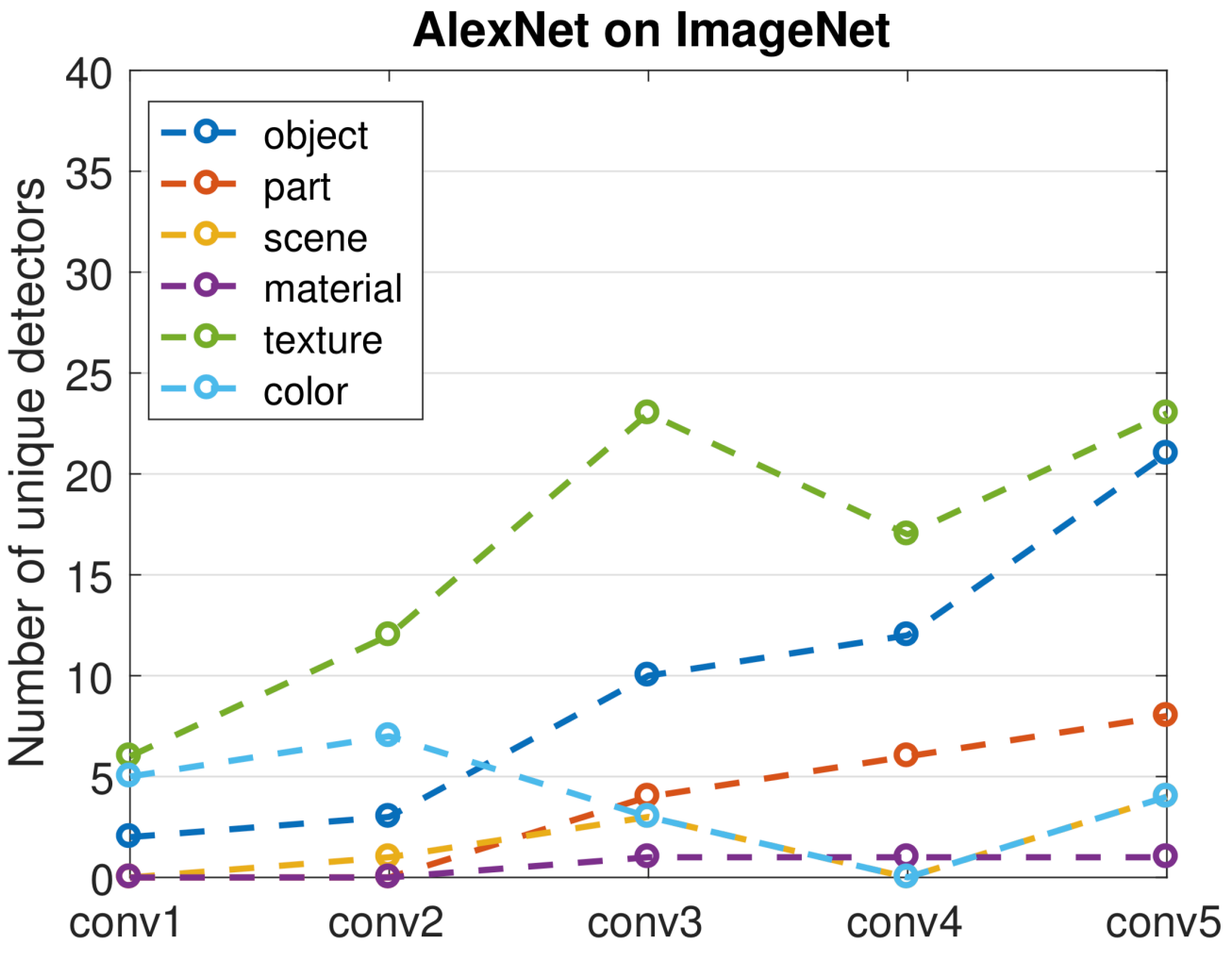
conv5 unit 107 road (object) IoU=0.15

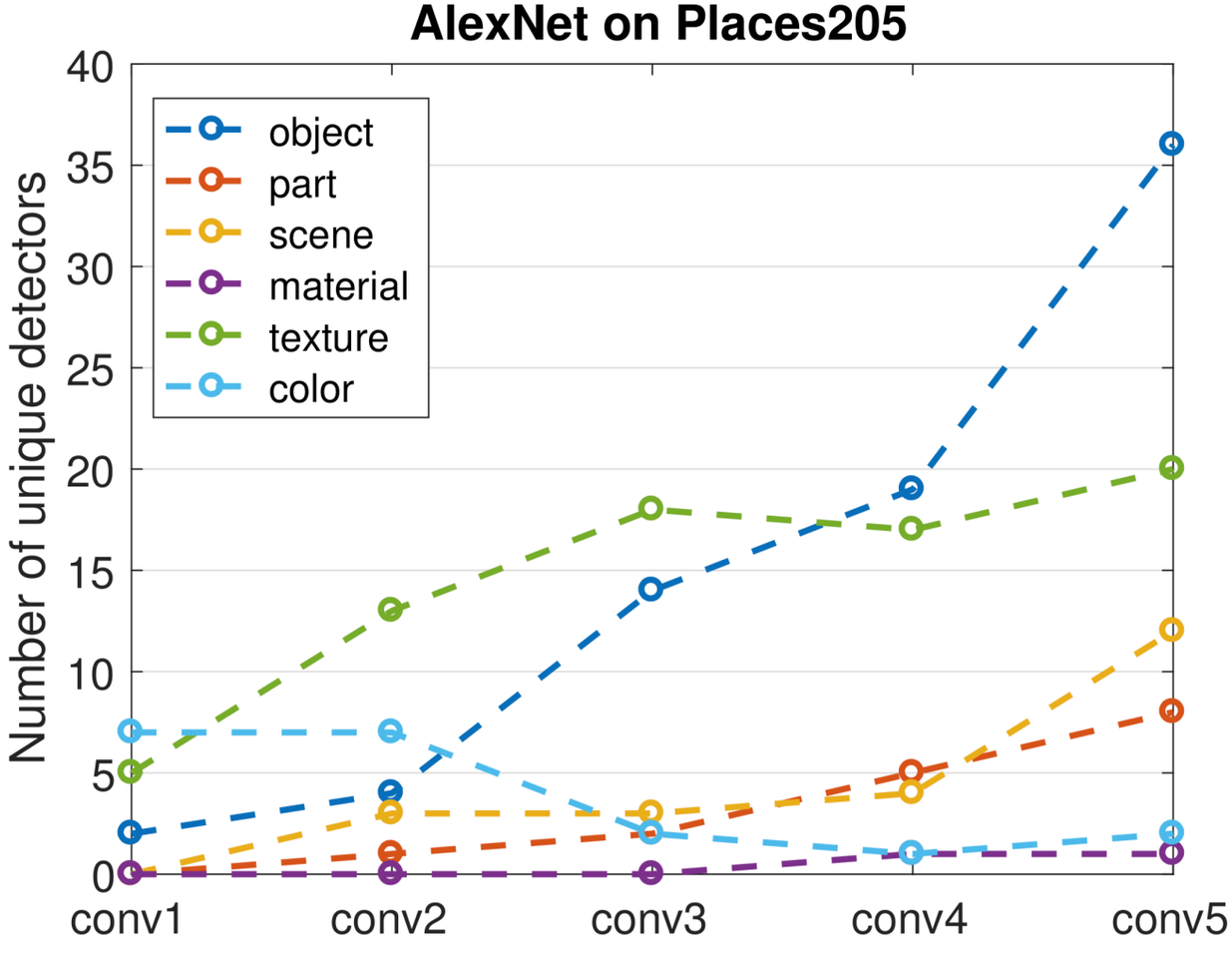
conv5 unit 79 car (object) IoU=0.13

Experiments

Quantifying interpretability of deep visual representations

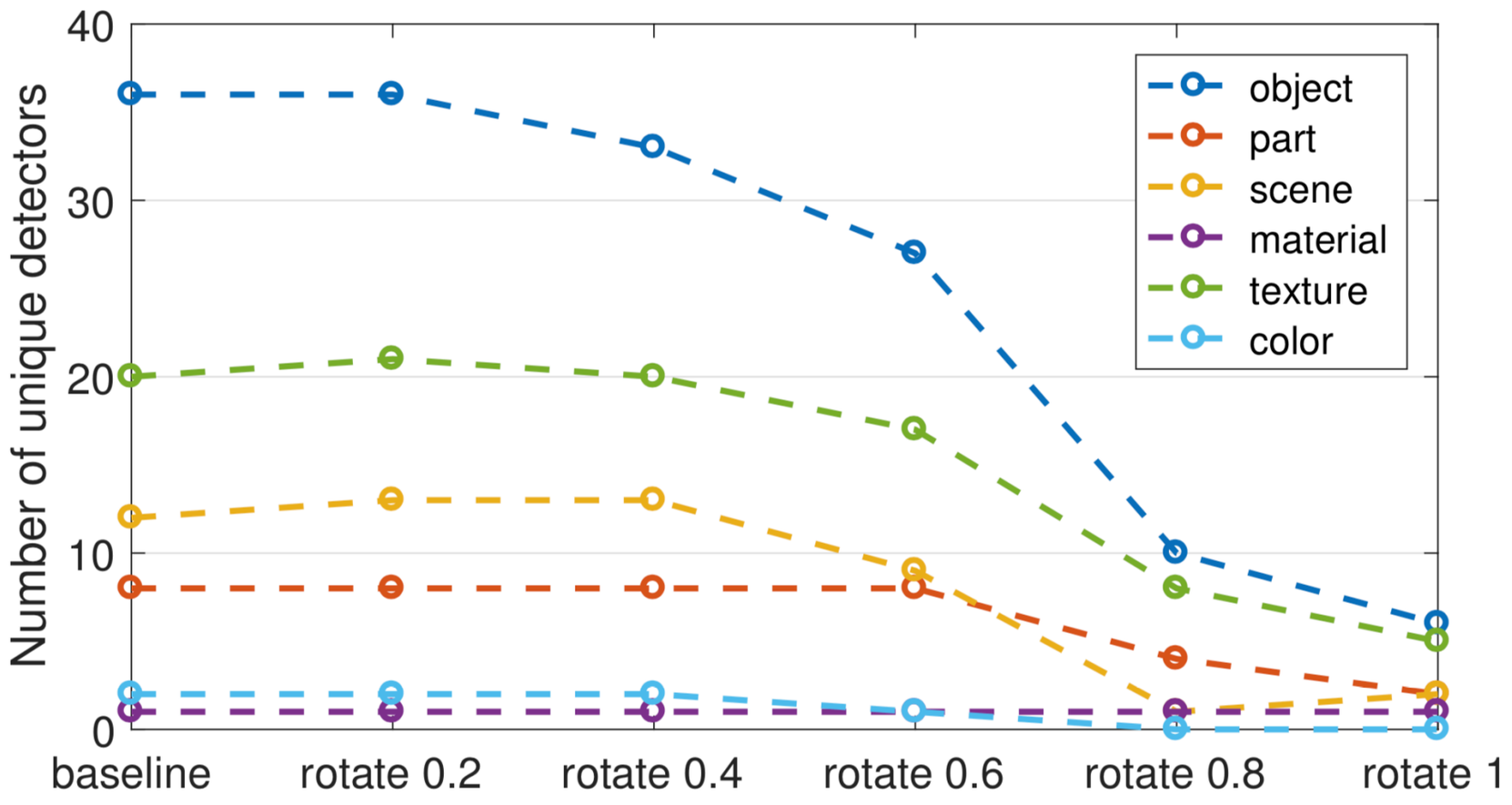
- Interpretability is quantified by how well the network aligns with a set of human interpretable concepts



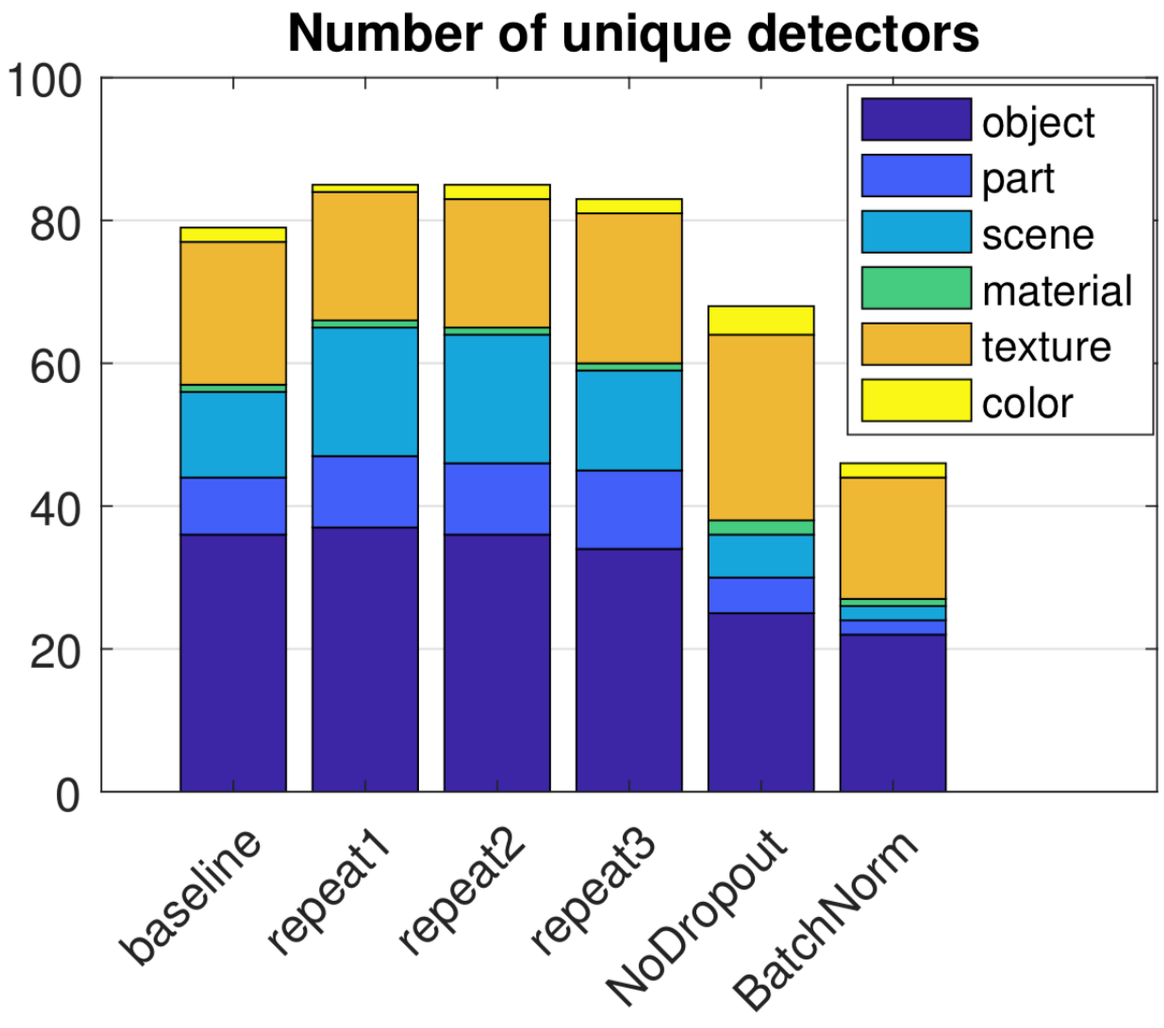


Interpretability != Discriminative Power

- Change the basis of the conv5 units in AlexNet to show that the interpretability can decrease while the discriminative power of the network stays constant



Effect of regularization on interpretability



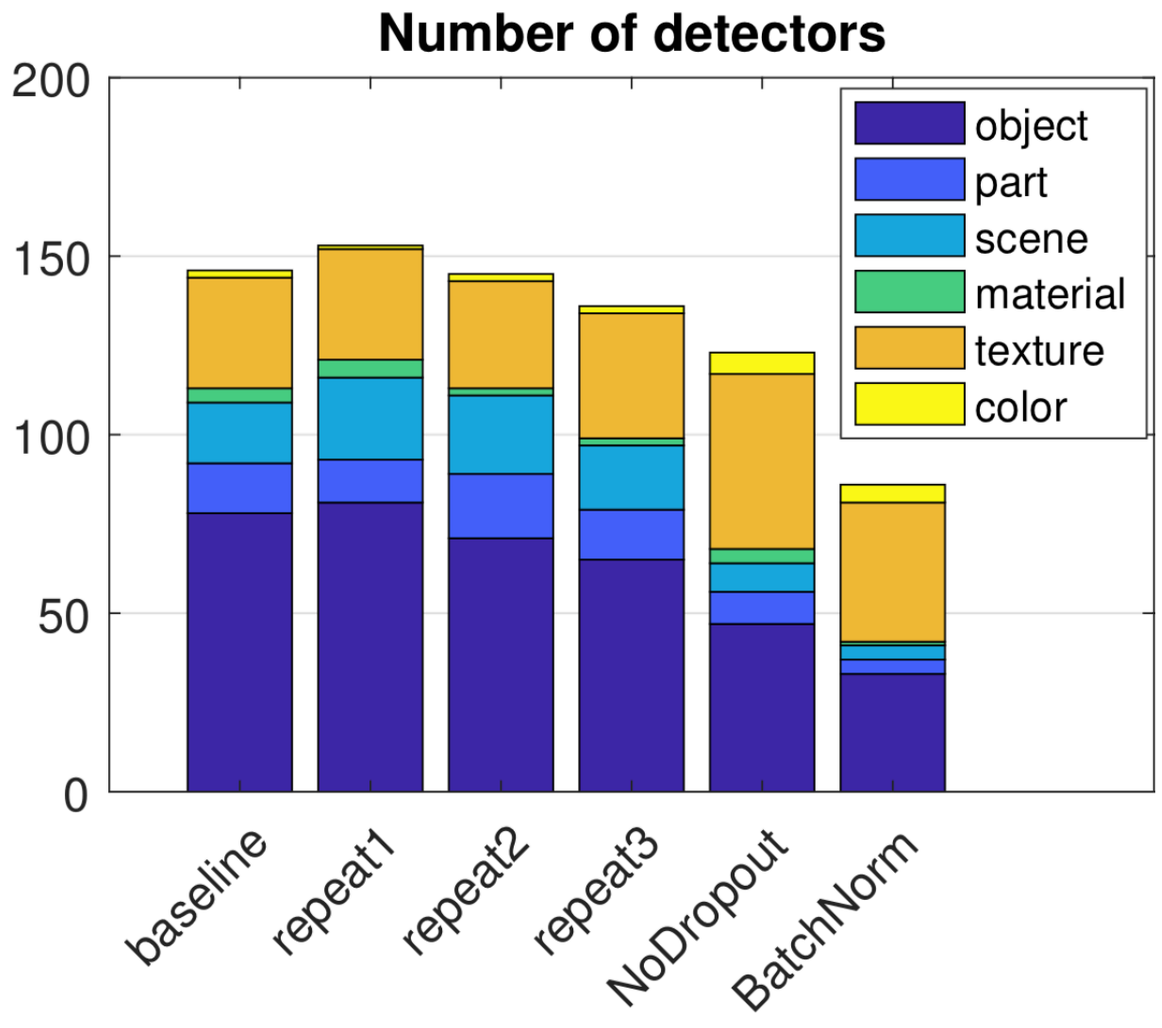
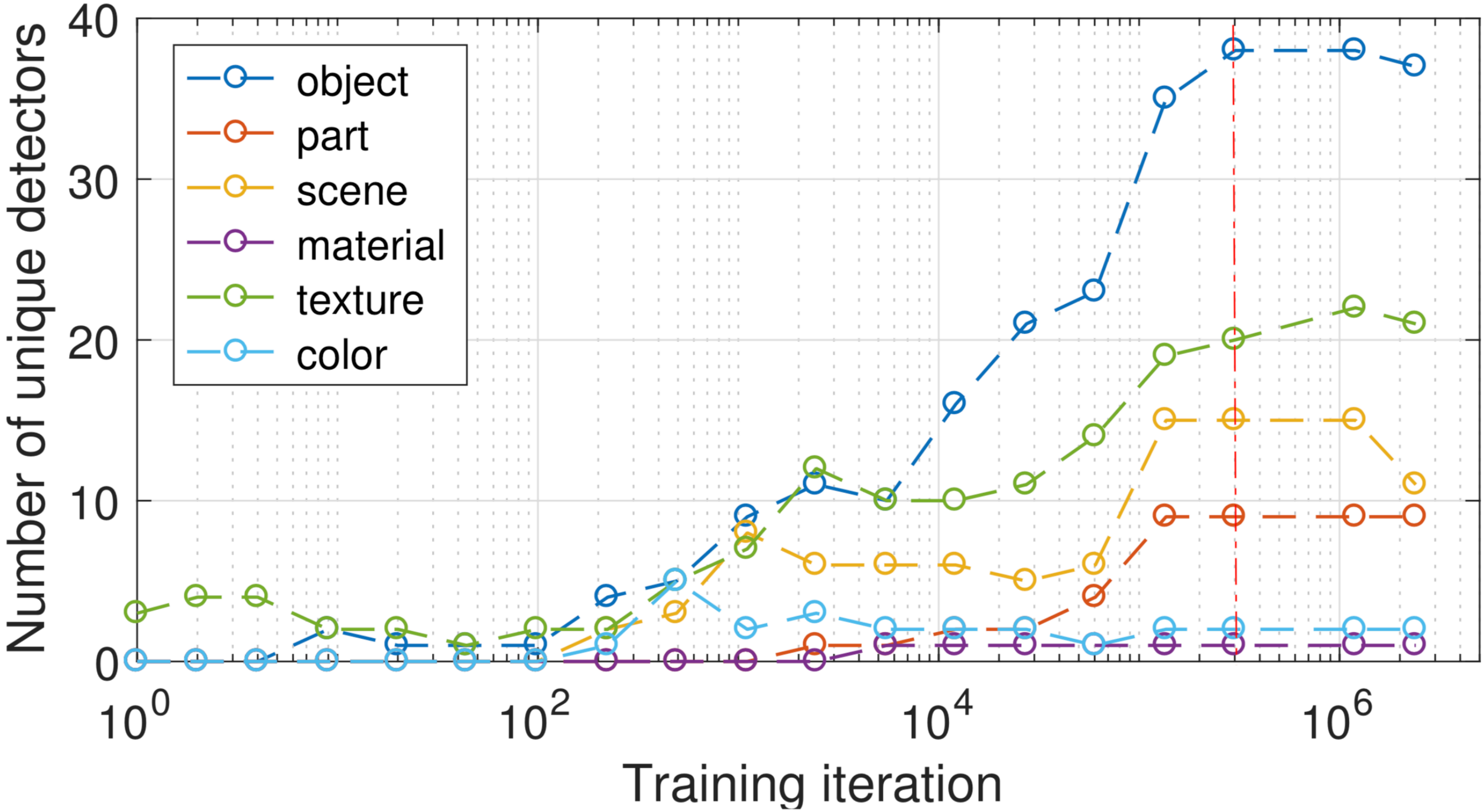


Figure 11. Effect of regularizations on the interpretability of CNNs.

Number of detectors vs epoch



Other experiments

- Random initialization does not seem to affect interpretability

- Widening of AlexNet showed an increase in the number of concept detectors

Thank you