OBJECT DETECTORS EMERGE IN DEEP SCENE CNN’S

Luca Simonetto - 11413522
Edgar Schönfeld - 11398272

Outline

- Introduction
- Methods and Results
    - Imagenet CNN and Places CNN
    - Internal CNN representation
    - Emergence of objects in the representation
- Conclusions
- Future work

Introduction - Paper Overview

Title: Object Detectors Emerge in Deep CNNs

Year: 2015

Concerned with: Understanding the internal representation learned by a CNN

Main Message: When training a CNN for *Scene Classification*,
Object detectors emerge as a byproduct.

Introduction - Paper Overview

Example of Scene Classification:

restaurant

street

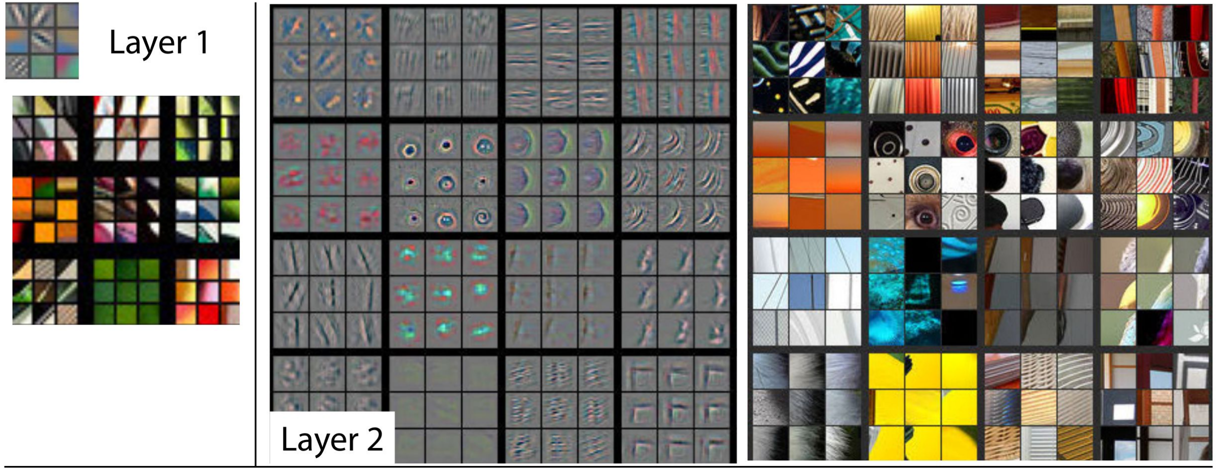
Additional Findings:

- CNN trained for *scene classification* naturally discovers **more** object categories than a CNN for *object recognition*

- The same network can do both *object localization* and *scene recognition* in a single forward-pass

Introduction - Related Work

- Visualizing and Understanding Convolutional Networks (2014)
- Analyzing the Performance of Multilayer Neural Networks for Object Recognition (2014)
- How transferable are features in deep neural networks? (2014)



Introduction - Scene vs. Object Classification

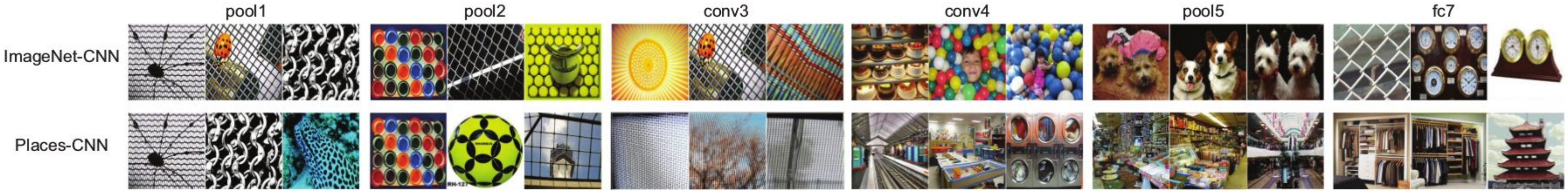
	Object Classification	**Scene Classification**
Constituents made of...	“Object Parts”	Objects
Parts have...	*strong* internal configuration	*weak* internal configuration
Consequence...	different and arbitrary part configurations	Less ambiguity
Object representation...	learned under supervision	Unsupervised learning

May explain why Scene-CNNs recognize objects so well

Methodology - Imagenet CNN and Places CNN

	ImageNet-CNN	Places-CNN
Network architecture:	Same	
Trained on images of...	Objects	Scenes
# of categories	1000	205
Top-1 accuracy	57.4 % - for object recog. 40.8 % - for scene recog. (with SVM)	50.0 %

clear bias



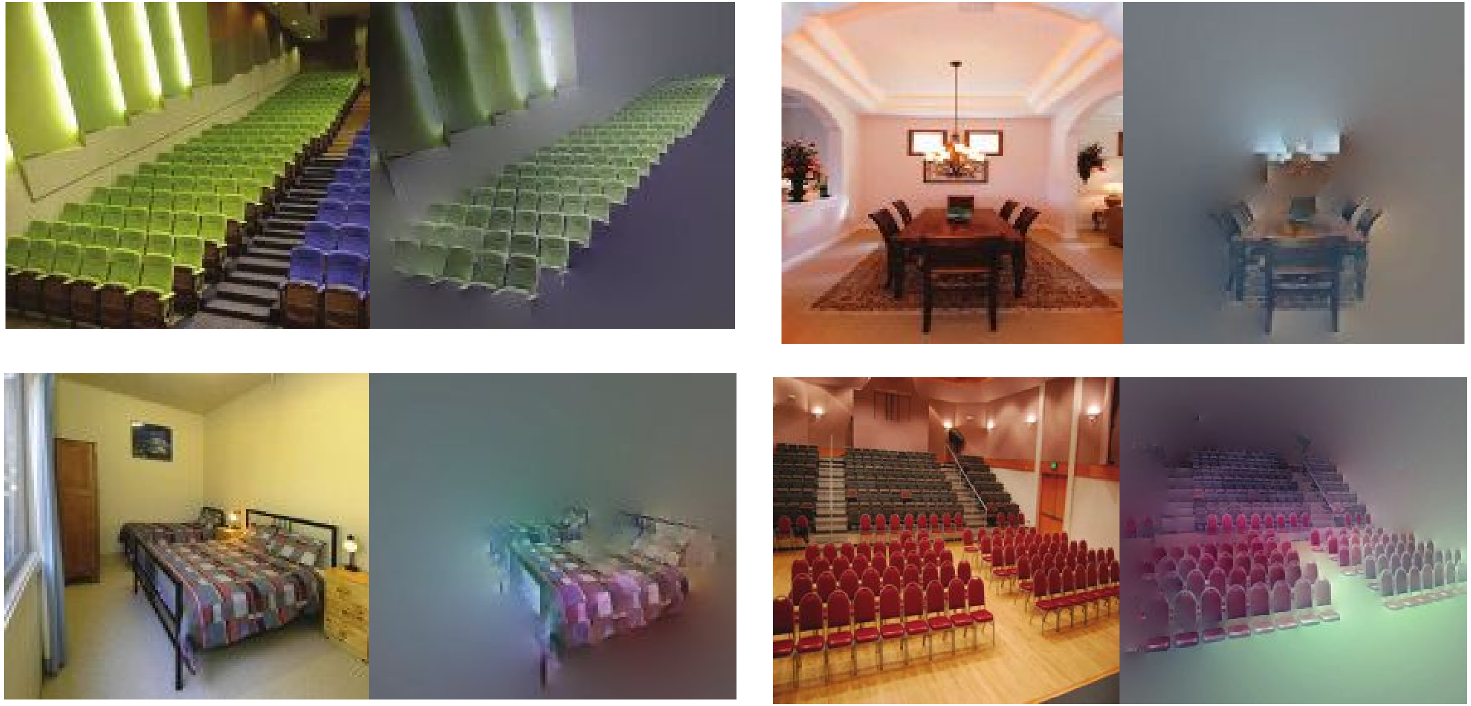
Internal CNN representation

How and what does a CNN learn?

Which parts of the image are used for classification?

Internal CNN representation

Simplifying the input representation - minimal images



Minimal Images

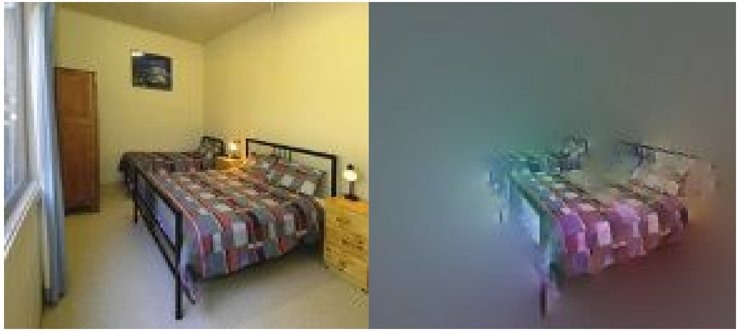
Classify image

↓

Segment image into regions

↓

Iteratively remove segments that
result in smallest decrease in classification score



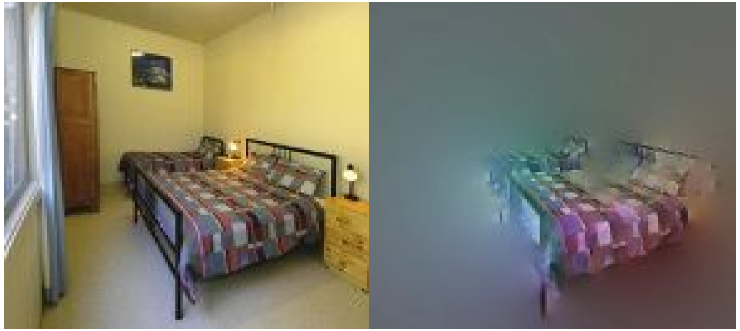
Minimal Images

Resulting image contains

minimal info to correctly classify image

mostly objects

Objects are important part of the representation built by the network



Bedroom - bed in 87% cases

Art gallery - paintings in 81% cases

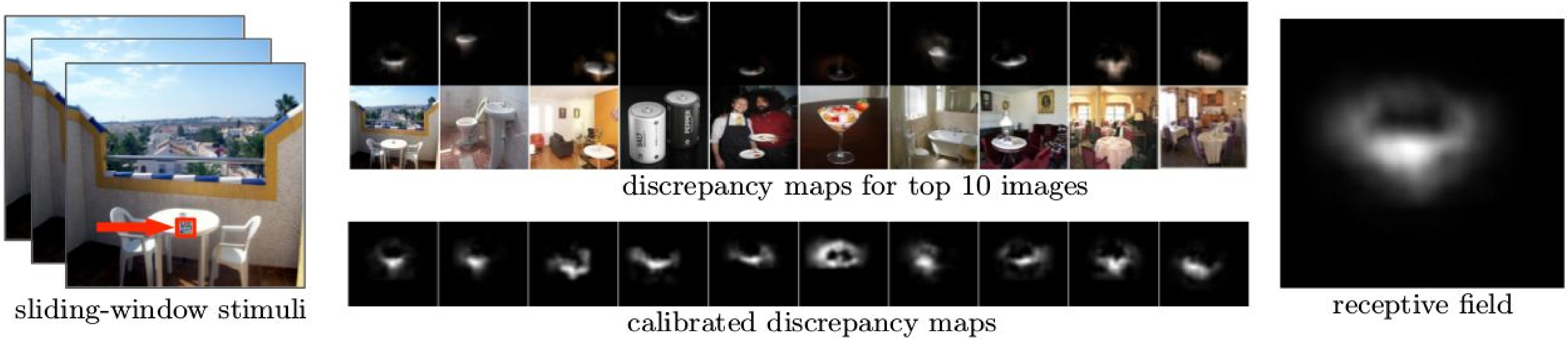
Bookstore - bookcase in 96% cases

Internal CNN representation

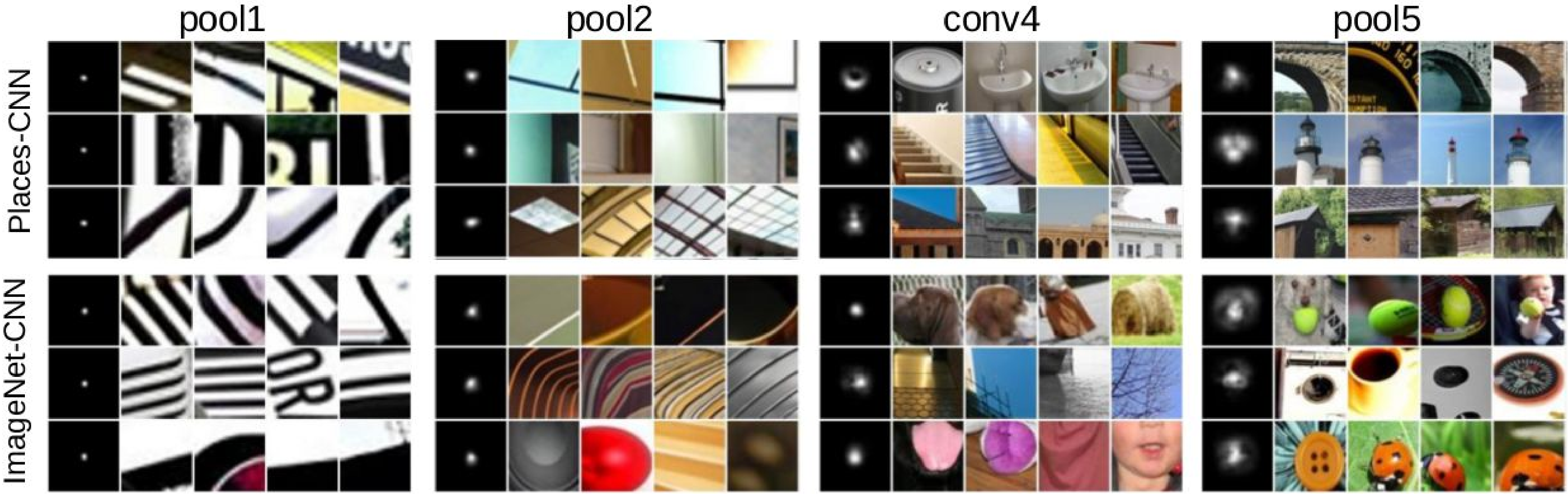
What is the shape and size of the internal receptive fields?

- Theoretical size defined by the CNN architecture
- Empirical size might be different

Determining the empirical rf size, using:



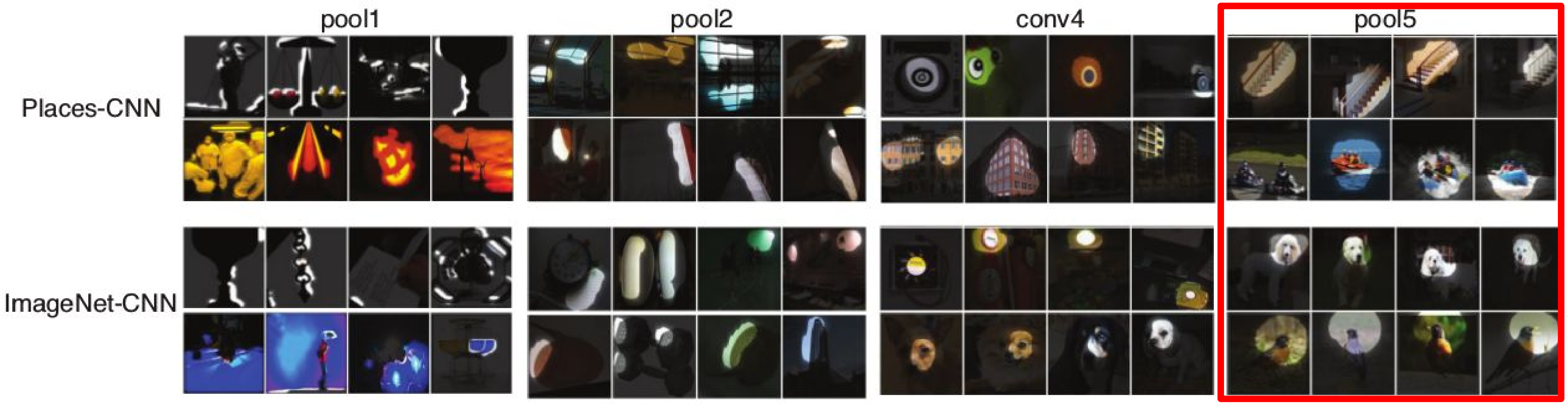
Internal CNN representation



The empirical size of the rf is much smaller than the theoretical size, especially in the later layers. They also are more meaningful deeper in the net.

Internal CNN representation

We can now do image segmentation (sort of), by looking at the receptive fields, and understand what the network “looks at” when doing scene classification.



Internal CNN representation

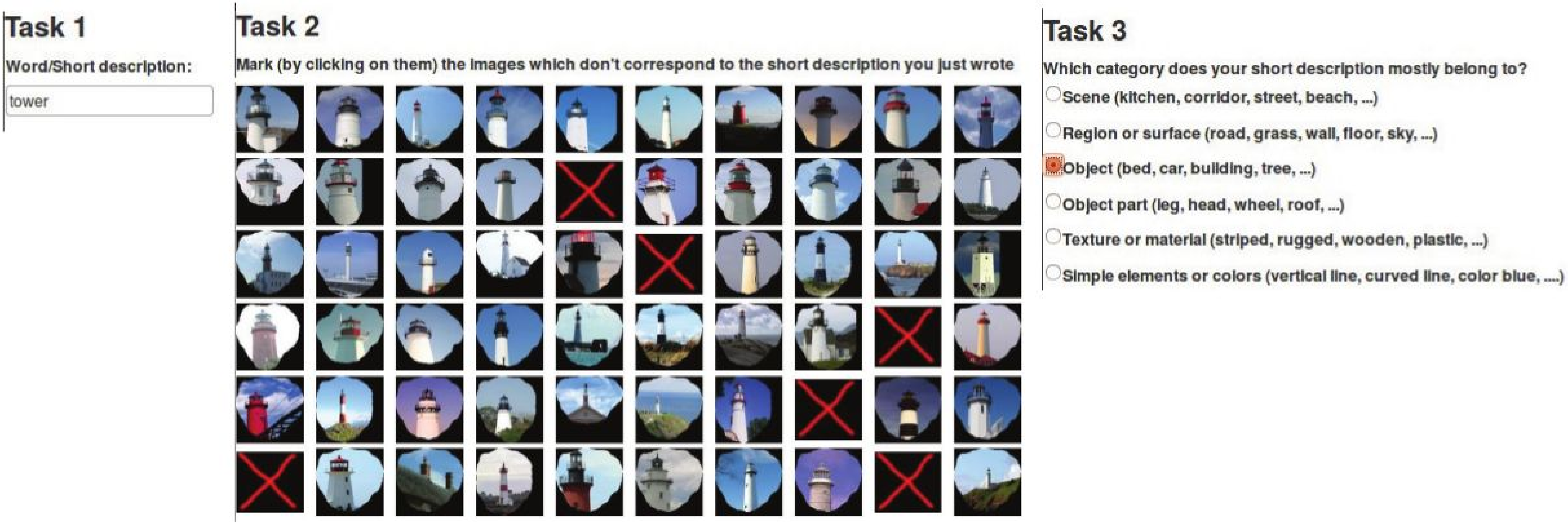
The deeper we go the more meaningful the rfs seems to be.

Do all the receptive fields work at the same abstraction level or not?

Some look for low level semantics (shapes, patterns) and others look for more complex ones (objects and scenes).

This analysis requires some brute force work → Amazon Mechanical Turk

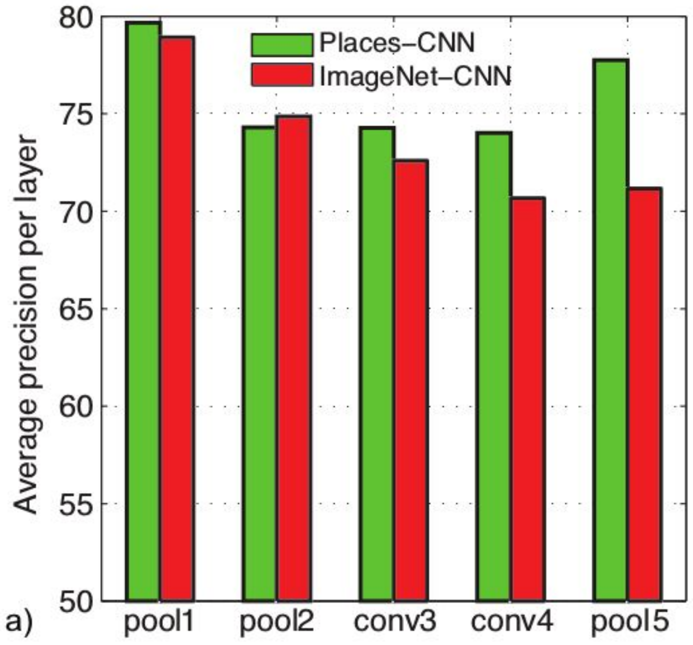
Internal CNN representation



Internal CNN representation

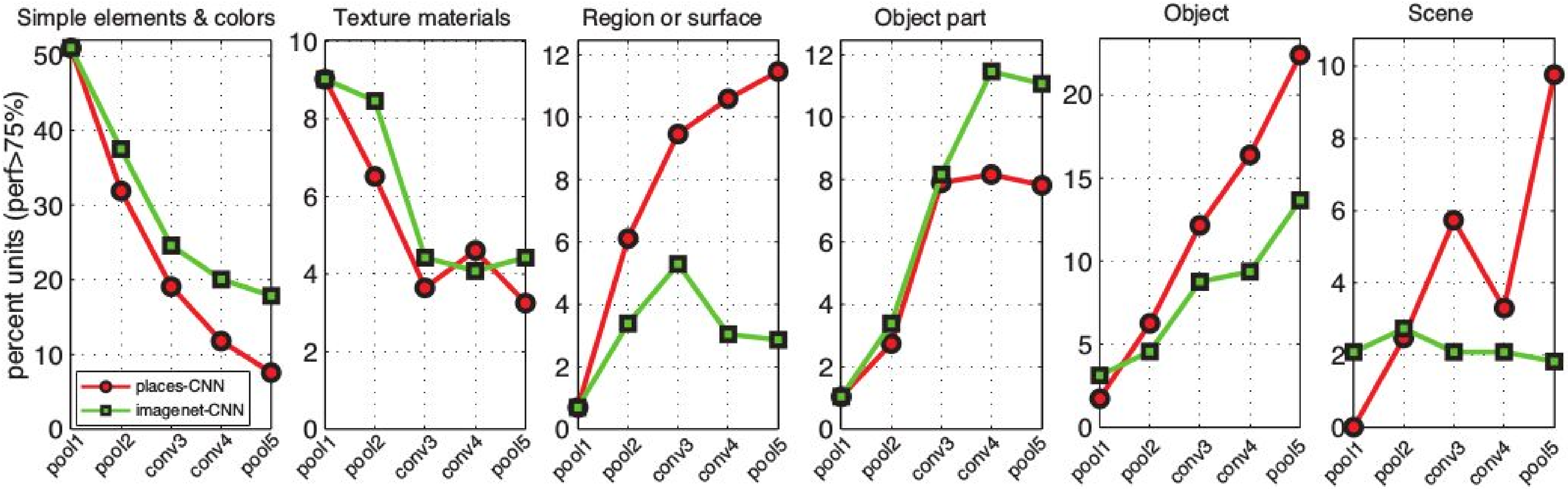
For each layer the average precision is calculated

Places-CNN layers usually have higher AP than imageNet-CNN!



Internal CNN representation

Distribution of concept categories (precision > 75%)



Emergence of objects in the representation

Deeper layers detect more high-level abstractions.

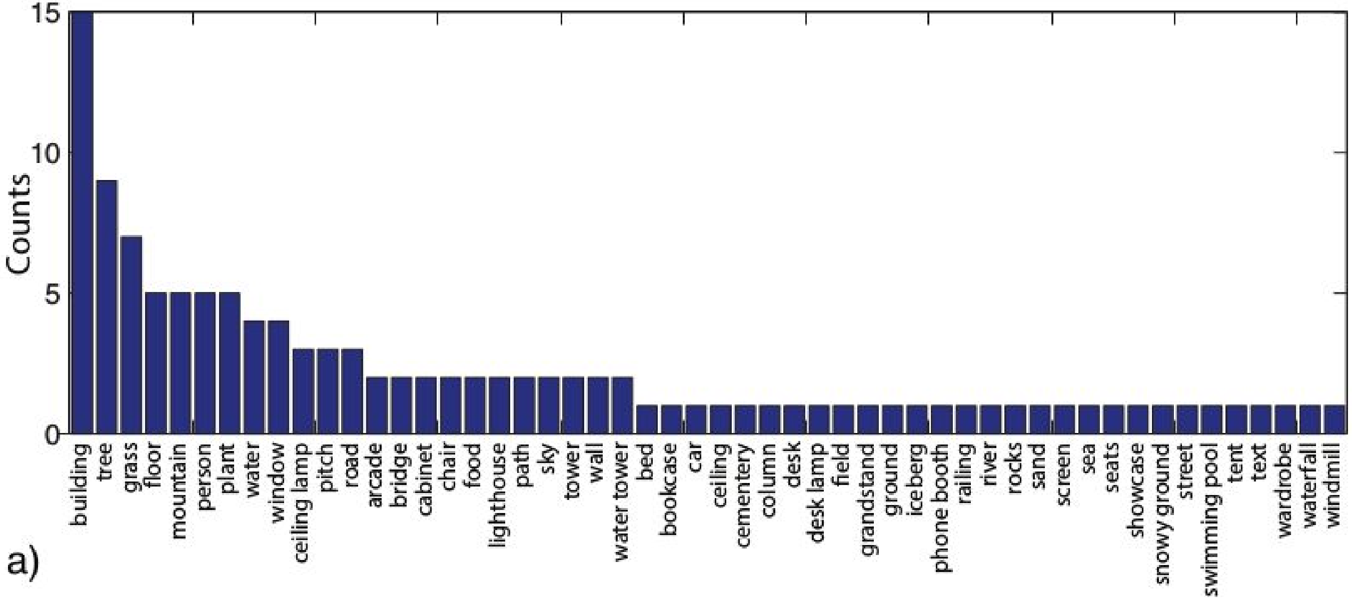
What object classes emerge?

Are multiple units detecting the same object?

Can we do segmentation with this information?

Emergence of objects in the representation

Objects detected in pool5 of Places-CNN

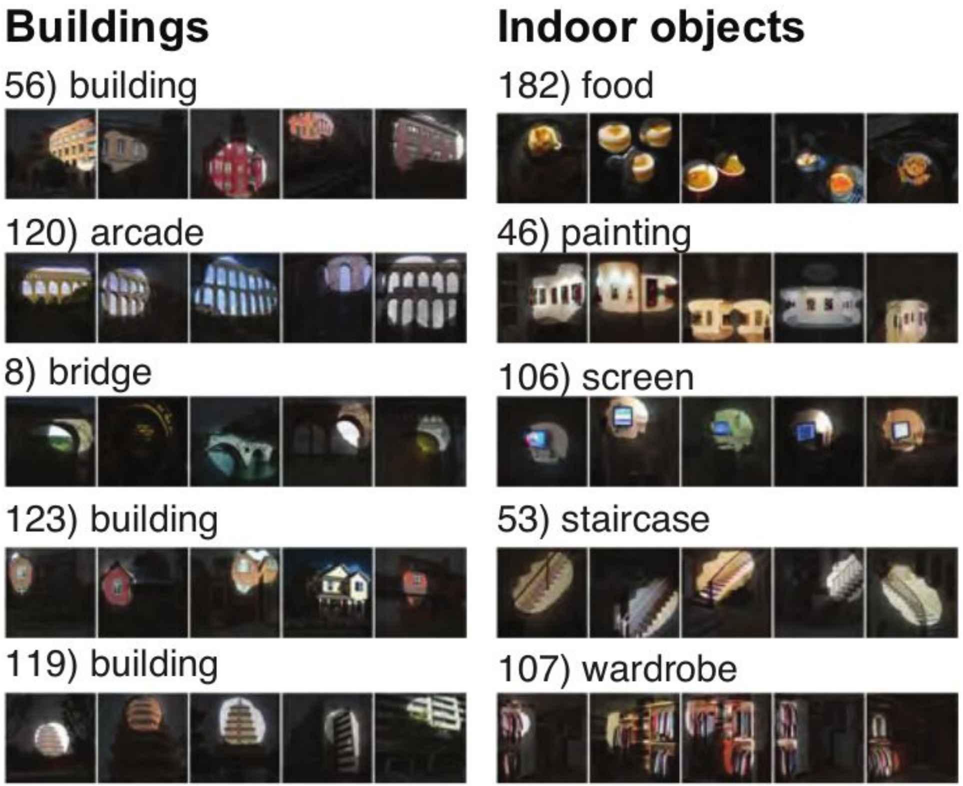


a)

Emergence of objects in the representation

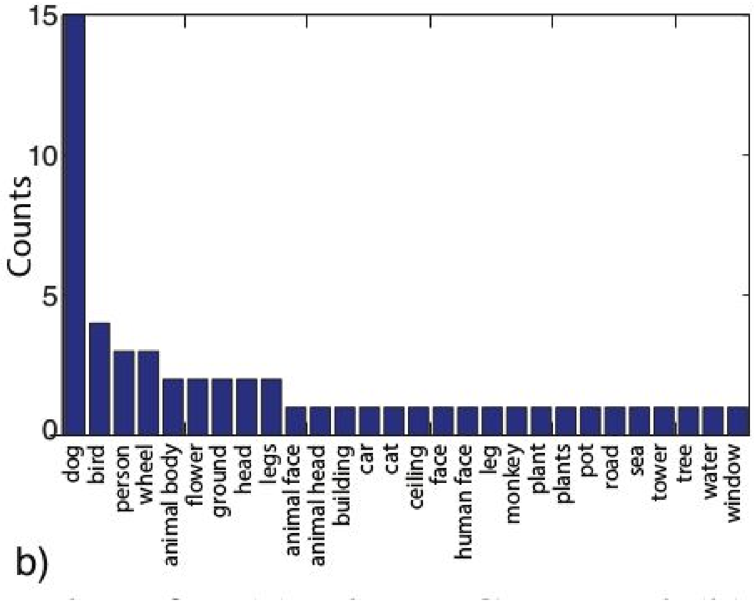
Many classes encoded by different units.

Each unit covers an object appearance.



Emergence of objects in the representation

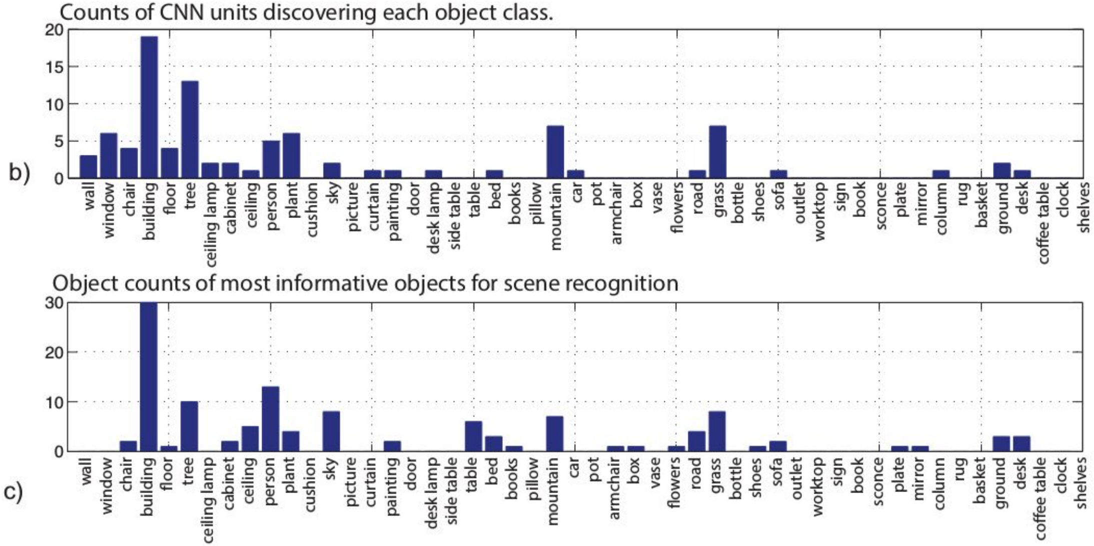
Objects detected in pool5 of ImageNet-CNN



Emergence of objects in the representation

Why do those object emerge?

- They seem to be correlated with the frequencies of the dataset used
- Objects detected seem to be the most informative for scene recognition

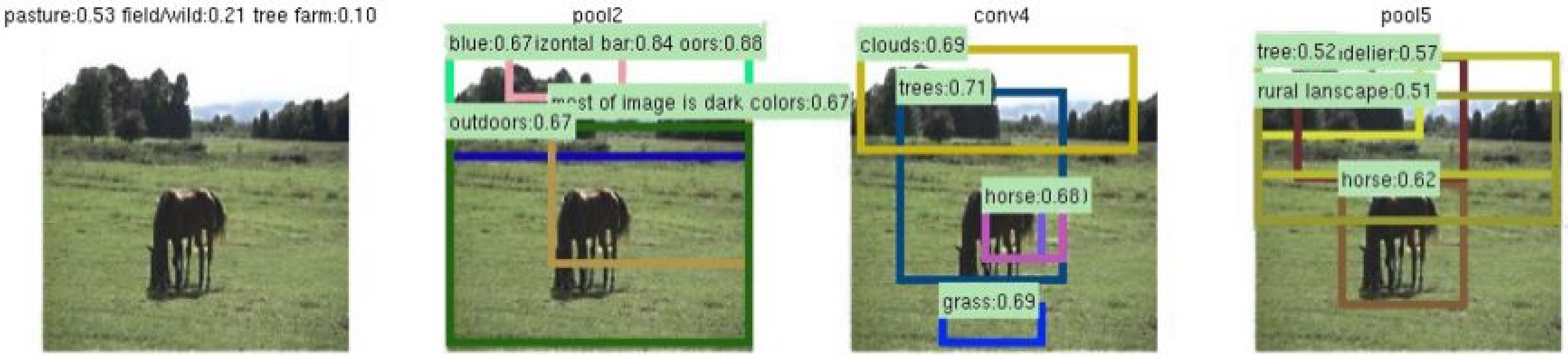


Emergence of objects in the representation

- Places-CNN achieves state-of-the-art performance on scene recognition
- The receptive fields are shaped around what they need to detect
- Every unit is specialized on a single concept
- Multiple units detect the same appearance of the same object
- Places-CNN detects the more discriminative objects

Let's do segmentation!

Emergence of objects in the representation



Conclusions

- CNNs that perform scene classification have developed internal object detectors
- Image segmentation can be done without being explicitly asked
- Scene recognition CNNs automatically learn which object are more discriminative than others

Future work

- Study of this phenomena, as it will probably appear in other CNNs for classification
- Constraints could be added to improve the internal quality of the receptive fields
- Develop an all-in-one system that can reliably combine the tasks exposed in this paper

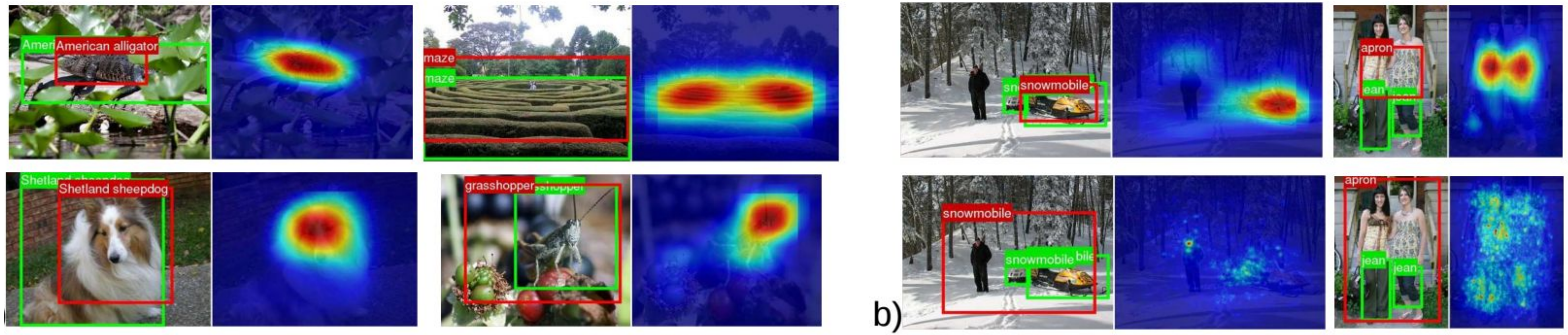


Figure: "Learning Deep Features for Discriminative Localization" (Zhou et al 2016)