Visualizing Deep Networks

Xiaolong Wang

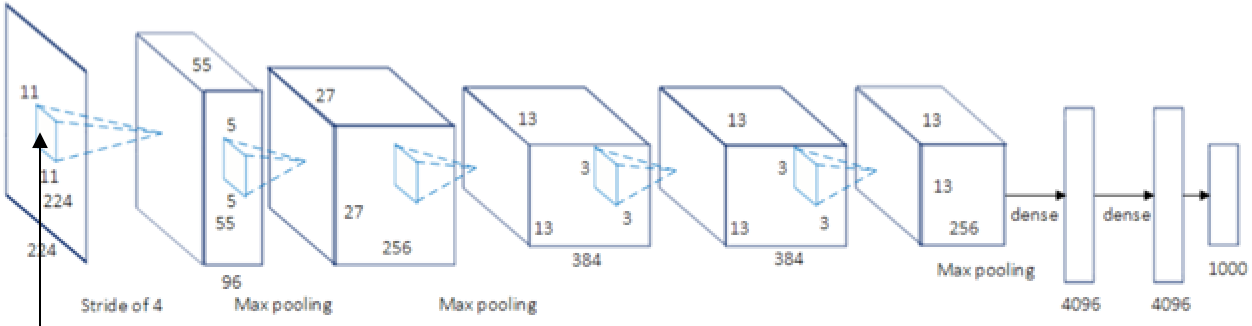
This Class: Visualizing Deep Networks

- Basics
- Visualizing “Saliency map”
- Visualization by maximizing activation
- Quantification on the units

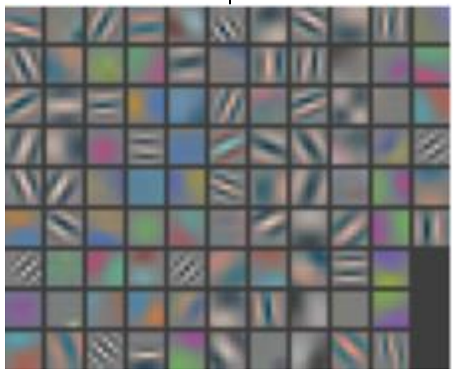
Slides partially from: http://cs231n.stanford.edu/ https://slazebni.cs.illinois.edu/fall20/

Basics

Basic visualization methods



Visualize first-layer weights directly



Basic visualization methods

layer 1 weights

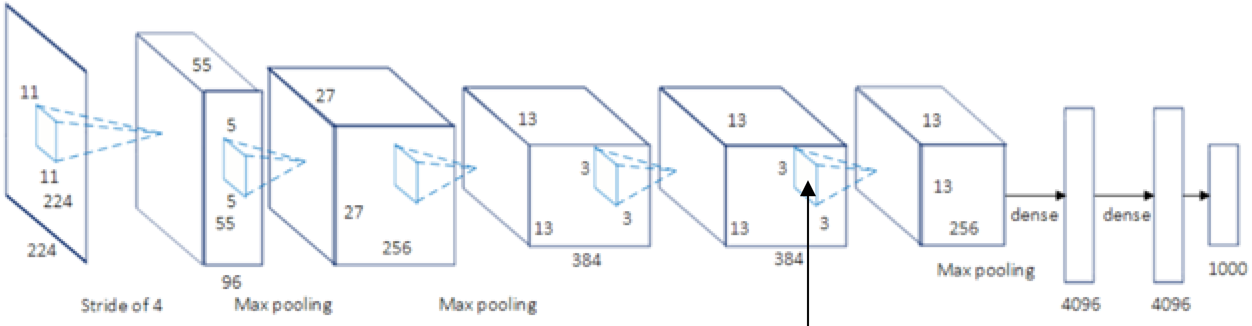
16 x 3 x 7 x 7

layer 2 weights

20 x 16 x 7 x 7

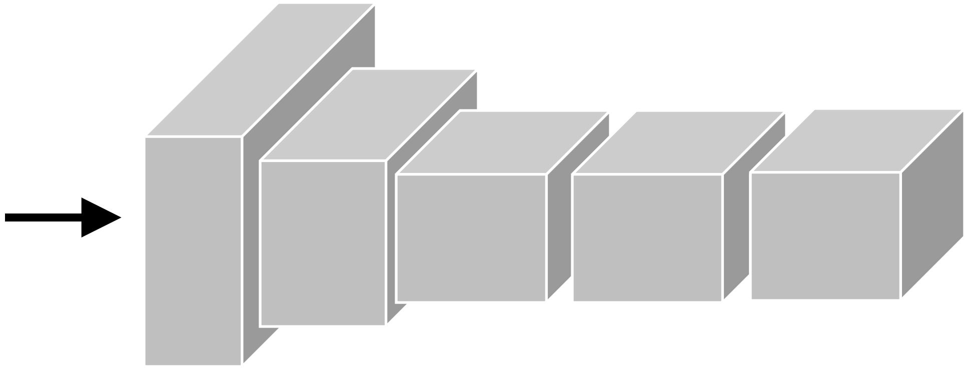
Visualizing the following layers is not very helpful

Basic visualization methods



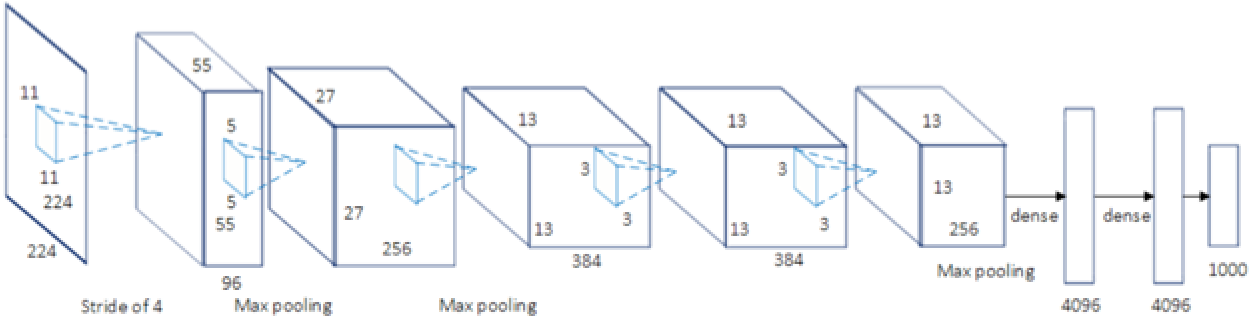
Visualize maximally activating patches

Visualize maximally activating patches

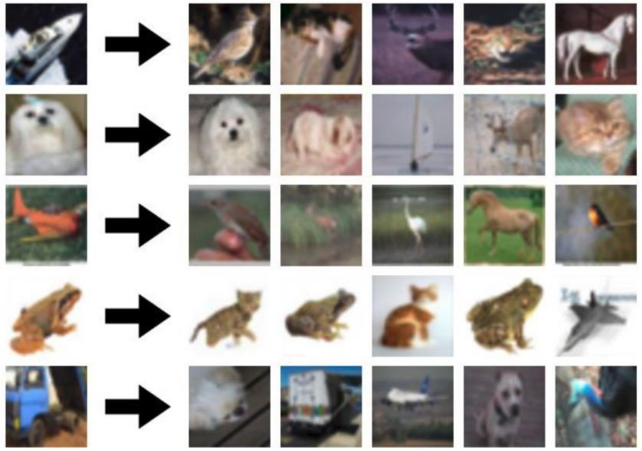


High response on one neuron

Basic visualization methods



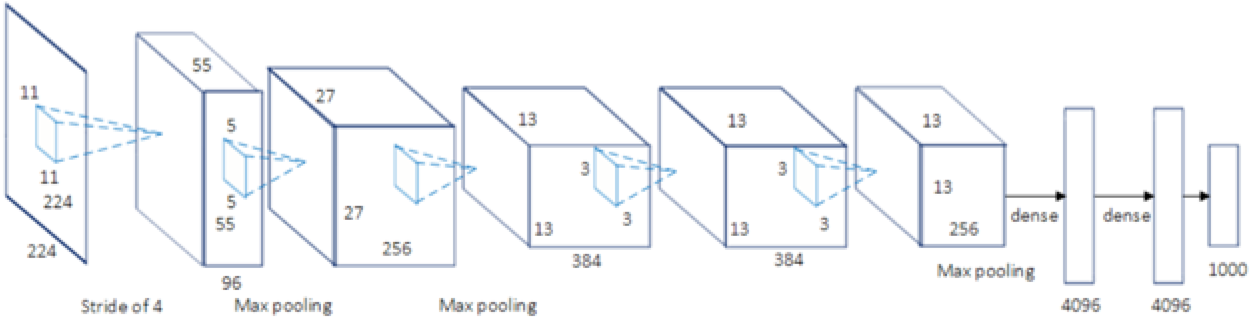
Recall: Nearest neighbors in <u>pixel</u> space



What about FC layers?

Visualize nearest neighbor images according to activation vectors

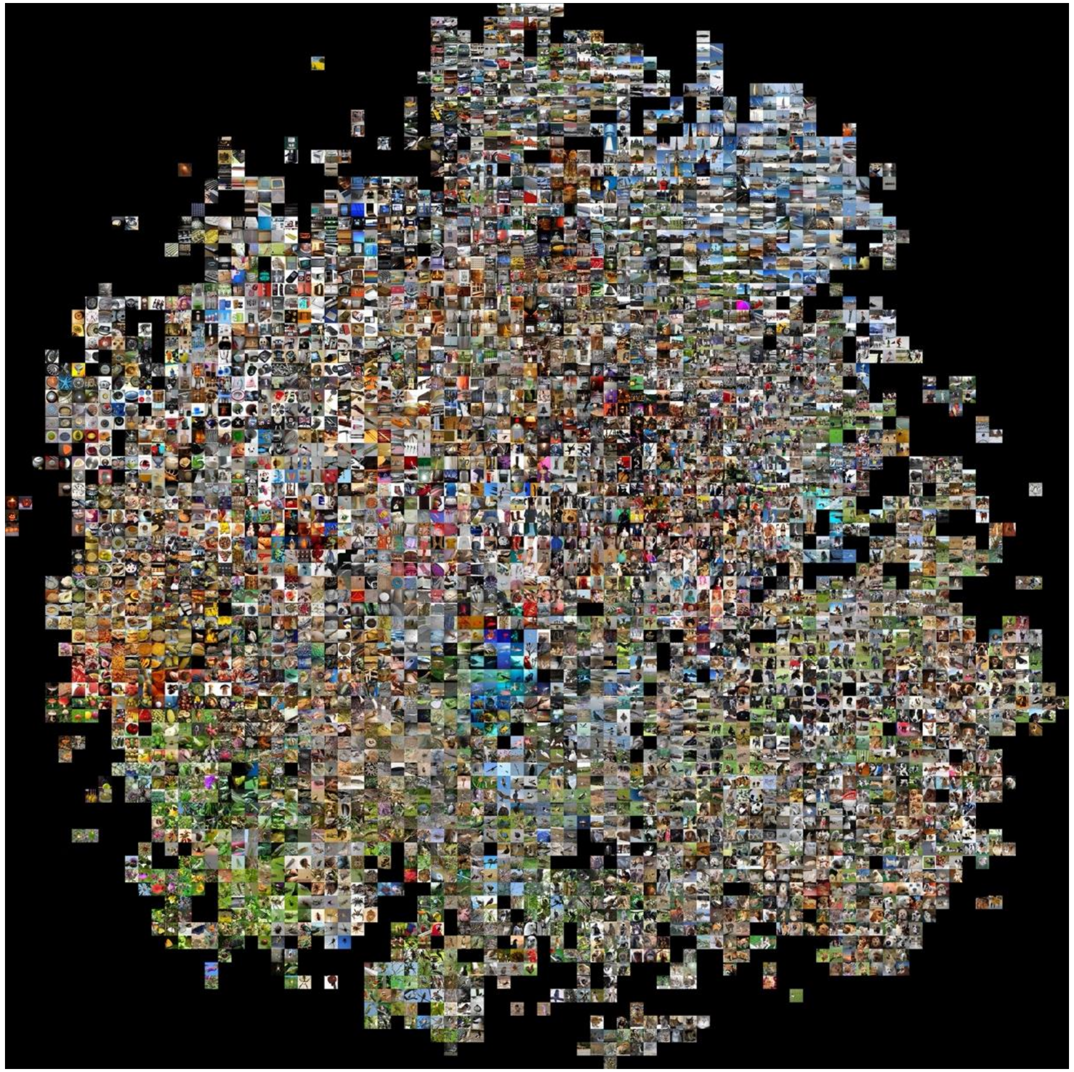
Basic visualization methods



What about FC layers?

dimensionality reduction, by t-SNE

Reduce image to 2-d using t-SNE



https://cs.stanford.edu/people/karpathy/cnnembed/cnn_embed_4k.jpg

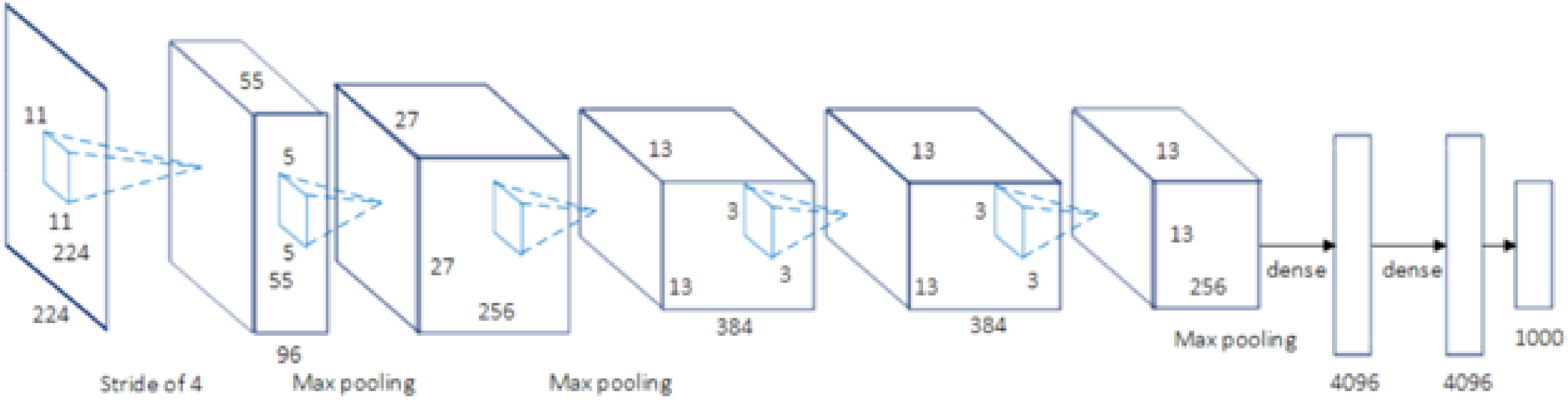
Visualizing “Saliency map”

Saliency map

Backpropagate gradient of class score (before softmax) to the image, display max of absolute values across color channels

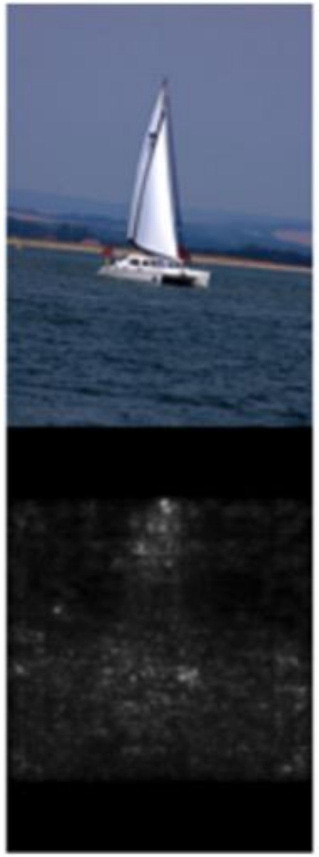
K. Simonyan, A. Vedaldi, and A. Zisserman, [Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps](), ICLR 2014

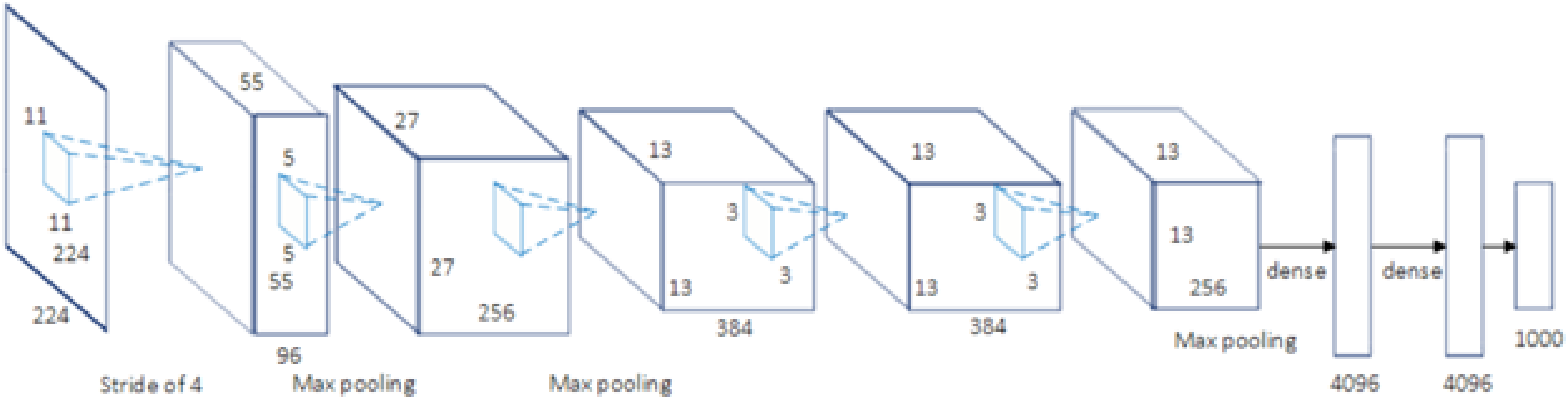
Saliency map



Compute the Loss

Saliency map



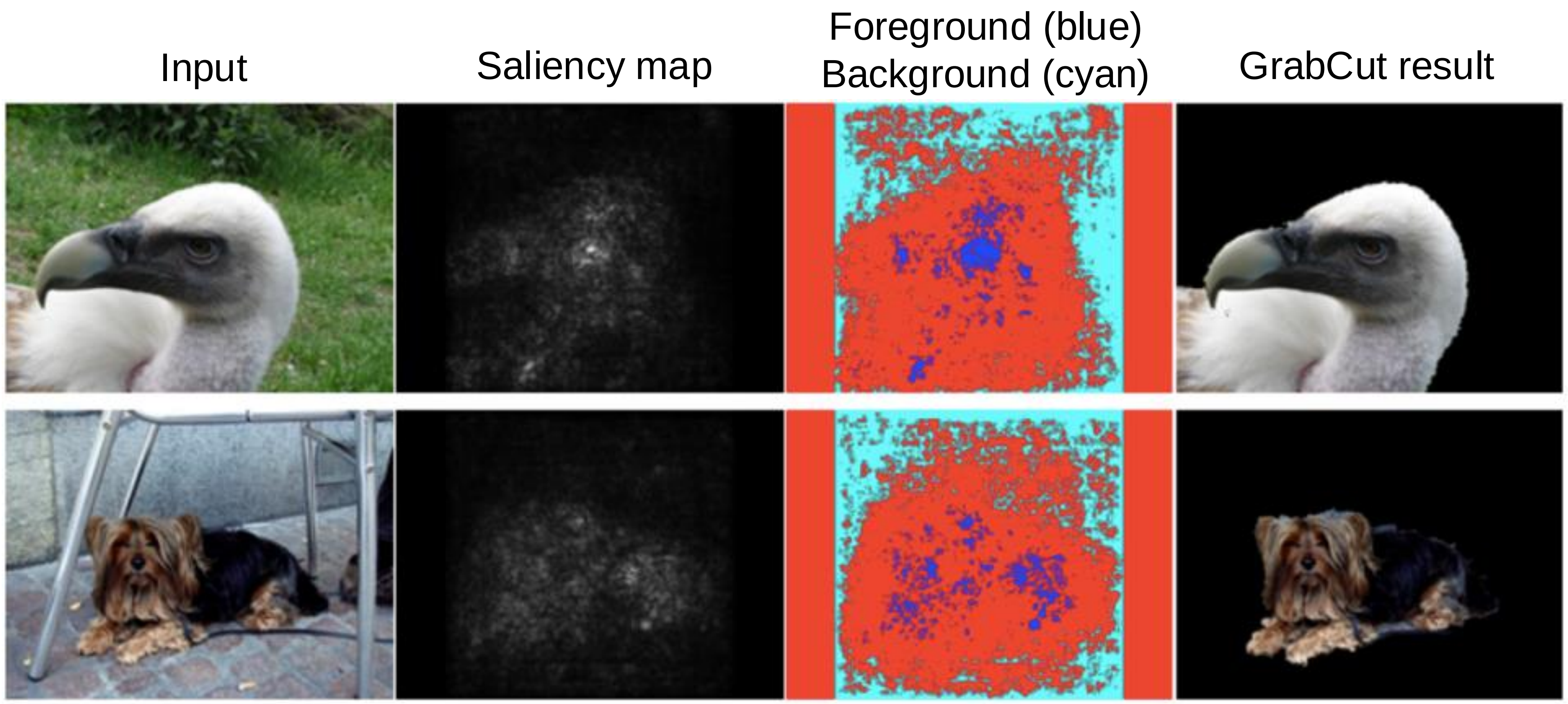


Compute the Loss

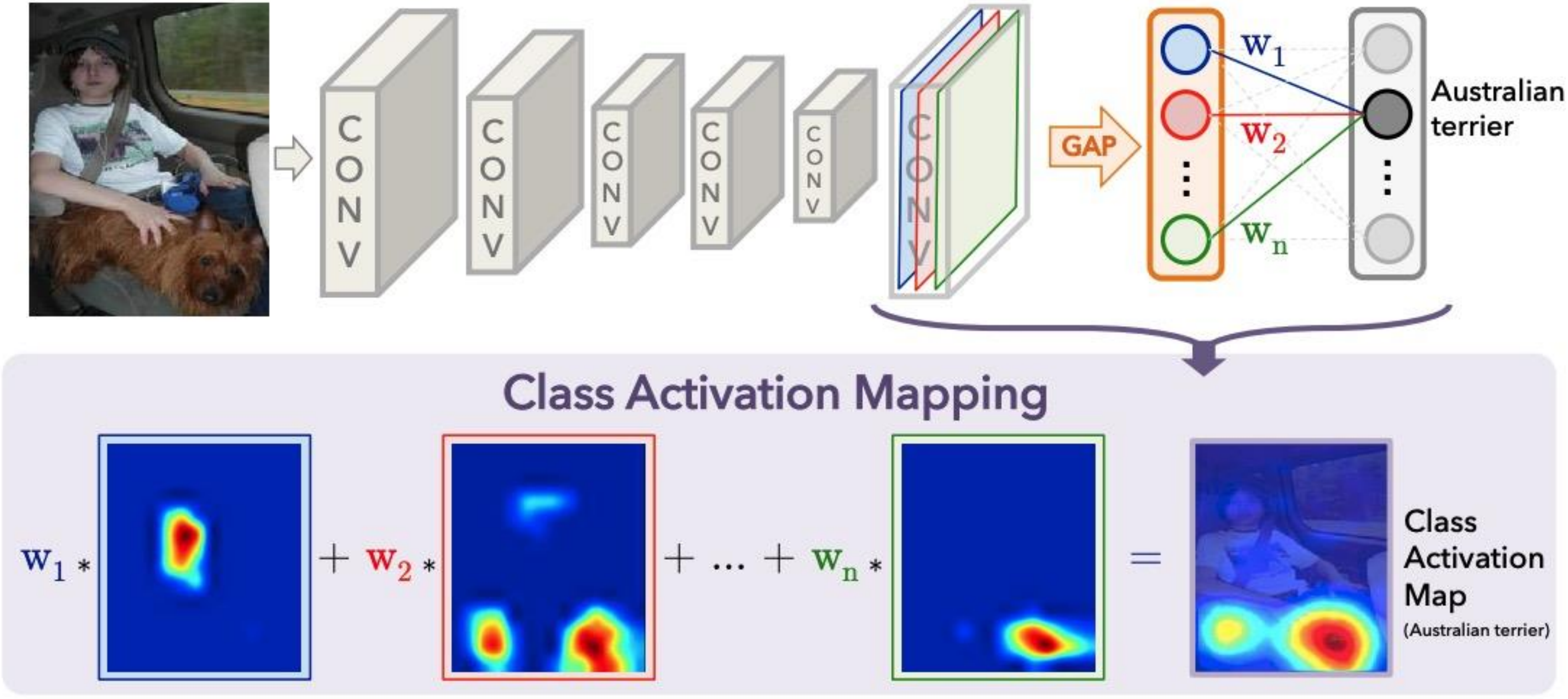
BP the gradients, but do not train the network
Visualize the final gradient to the image

Saliency map

Can be used for *weakly supervised* segmentation:



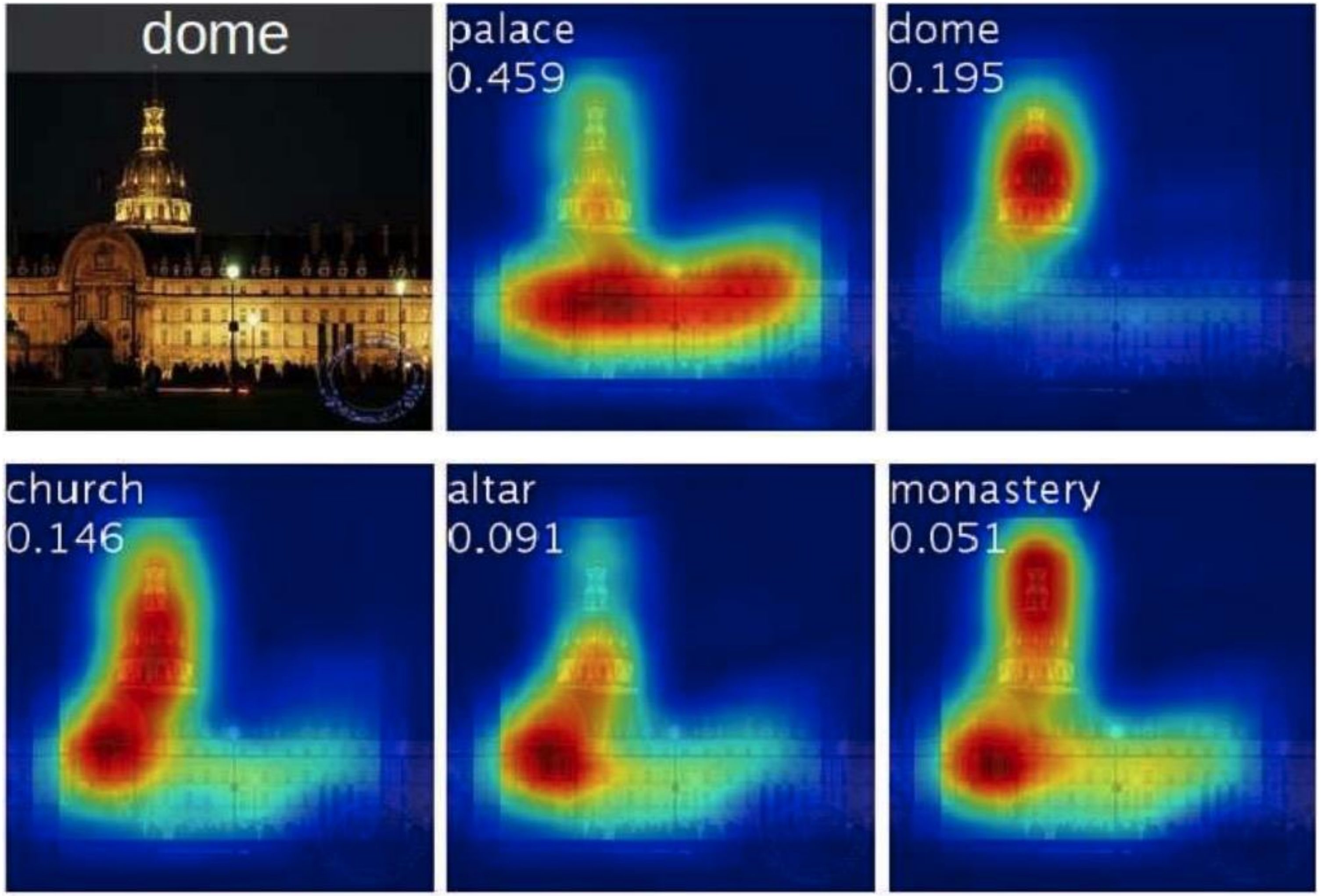
CAM: Class Activation Maps



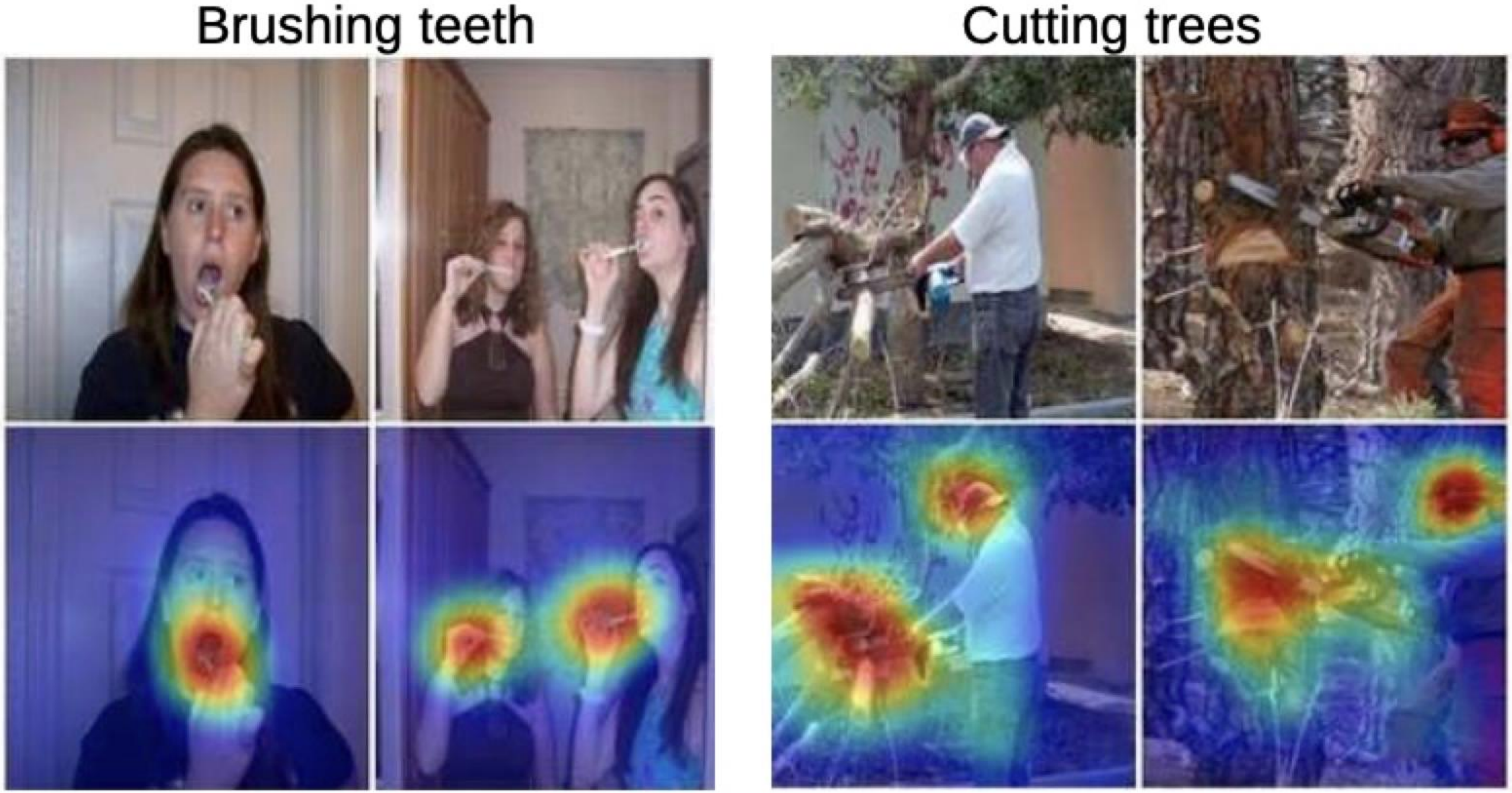
Zhou et al. 2016

CAM: Class Activation Maps

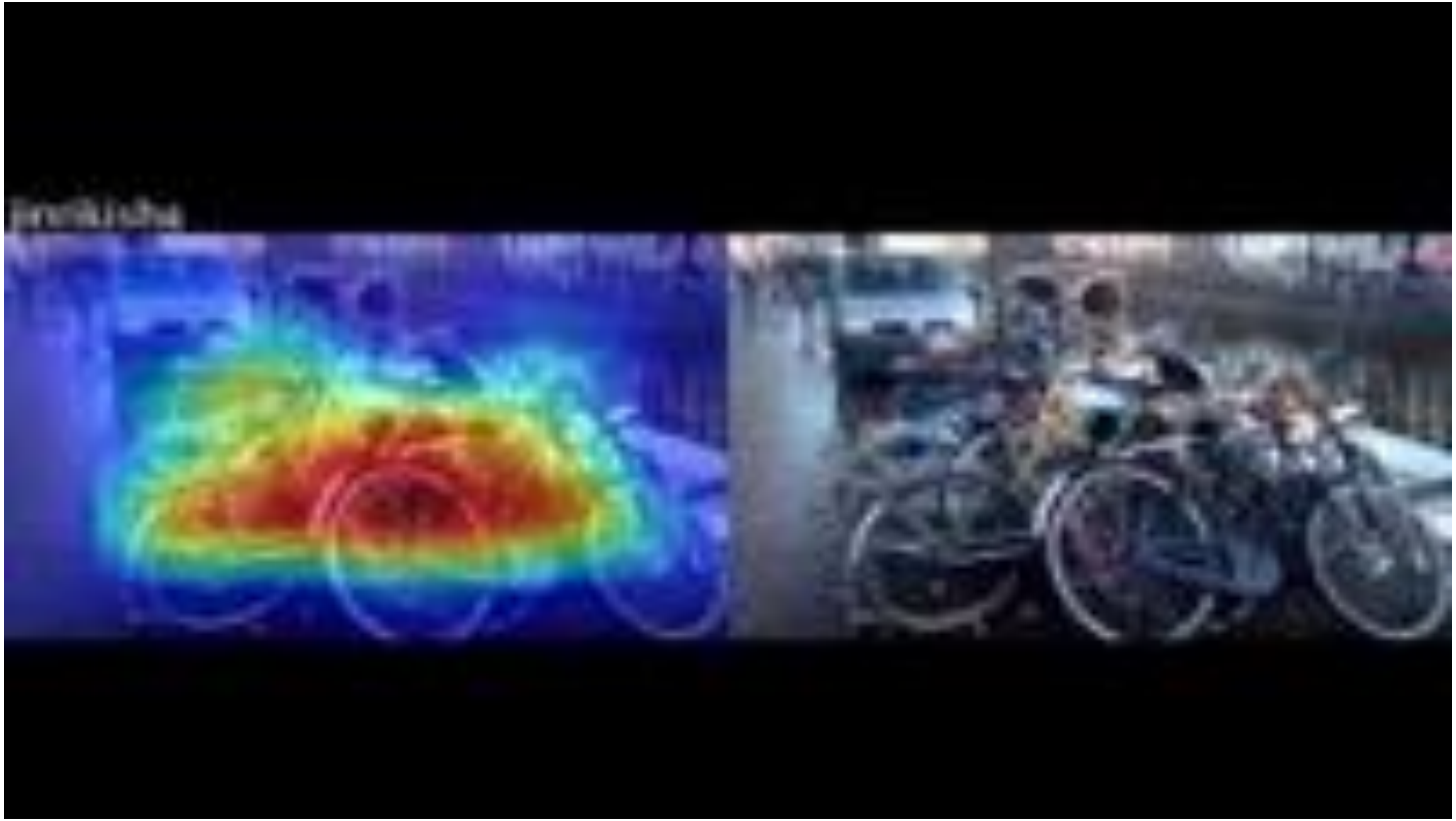
CAM: Class Activation Maps



CAM: Class Activation Maps



CAM: Class Activation Maps



Visualization by maximizing activation

Visualization by optimization

- We can synthesize images that maximize activation of a given neuron.

- Find image x maximizing target activation $f(x)$ subject to *natural image regularization penalty* $R(x)$:

$$x^* = \arg\max_x f(x) - \lambda R(x)$$

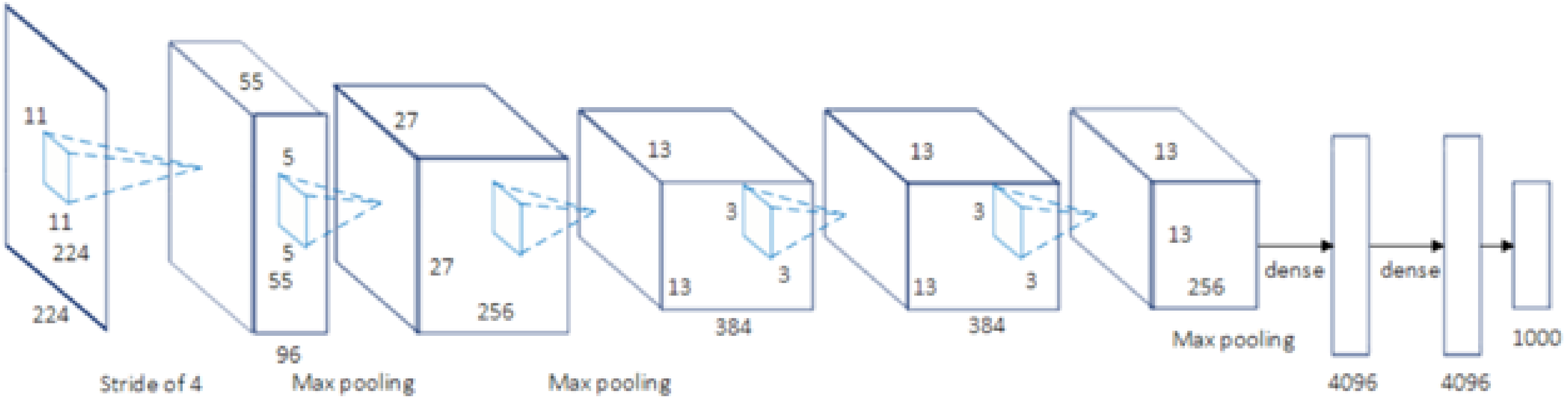
Visualization by optimization

- Maximize $f(x) - \lambda R(x)$
  - $f(x)$ is score for a category *before softmax*
  - $R(x)$ is L2 regularization
  - Perform *gradient ascent* starting with zero image, add dataset mean to result

dumbbell cup dalmatian

Simonyan et al. 2014

Visualization by optimization

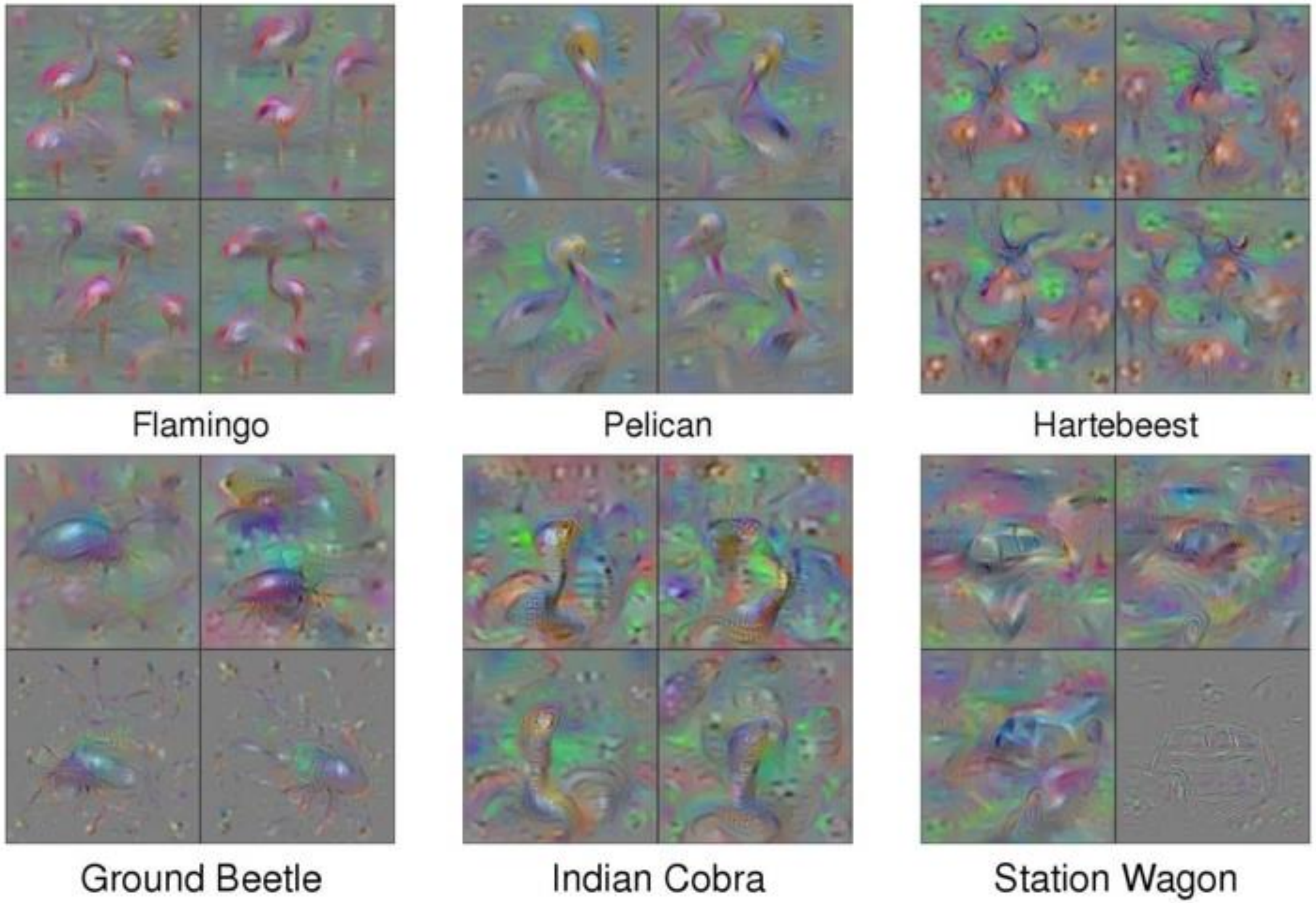


Compute the Loss

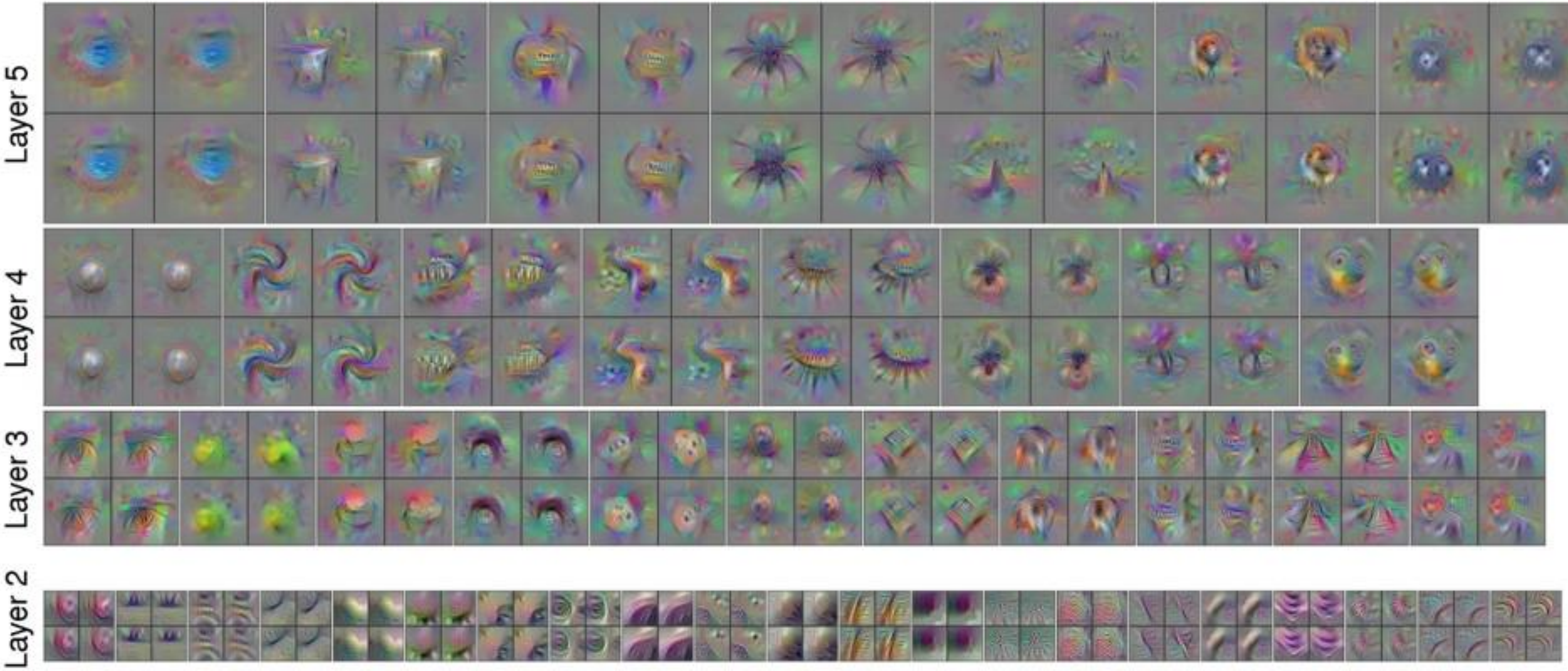
BP the gradients, but do not train the network

Keep adding/aggregating the gradients under constraint $R(x)$

Visualization by optimization

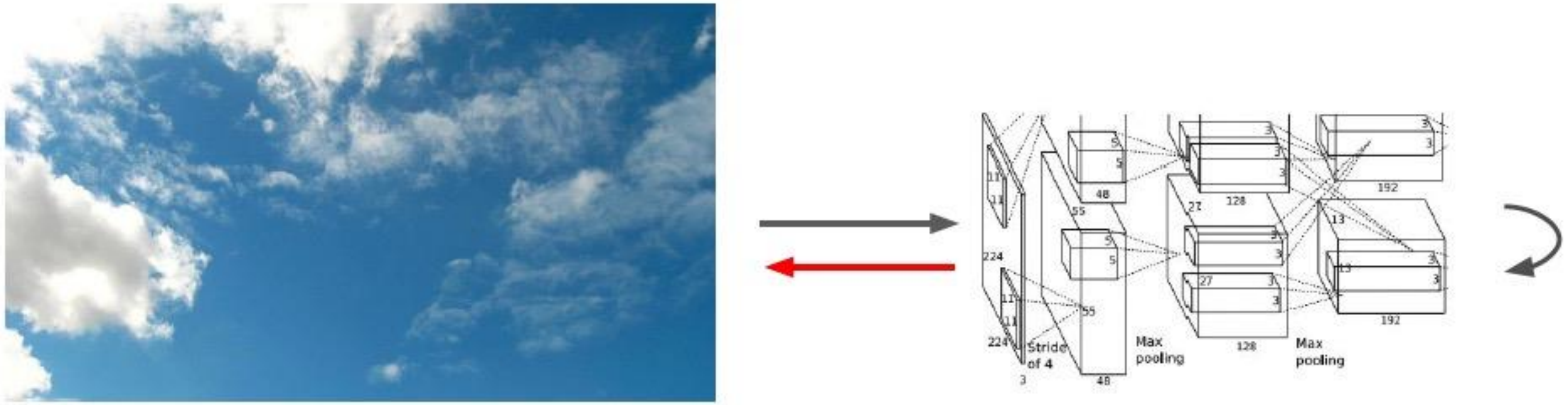


Visualization by optimization



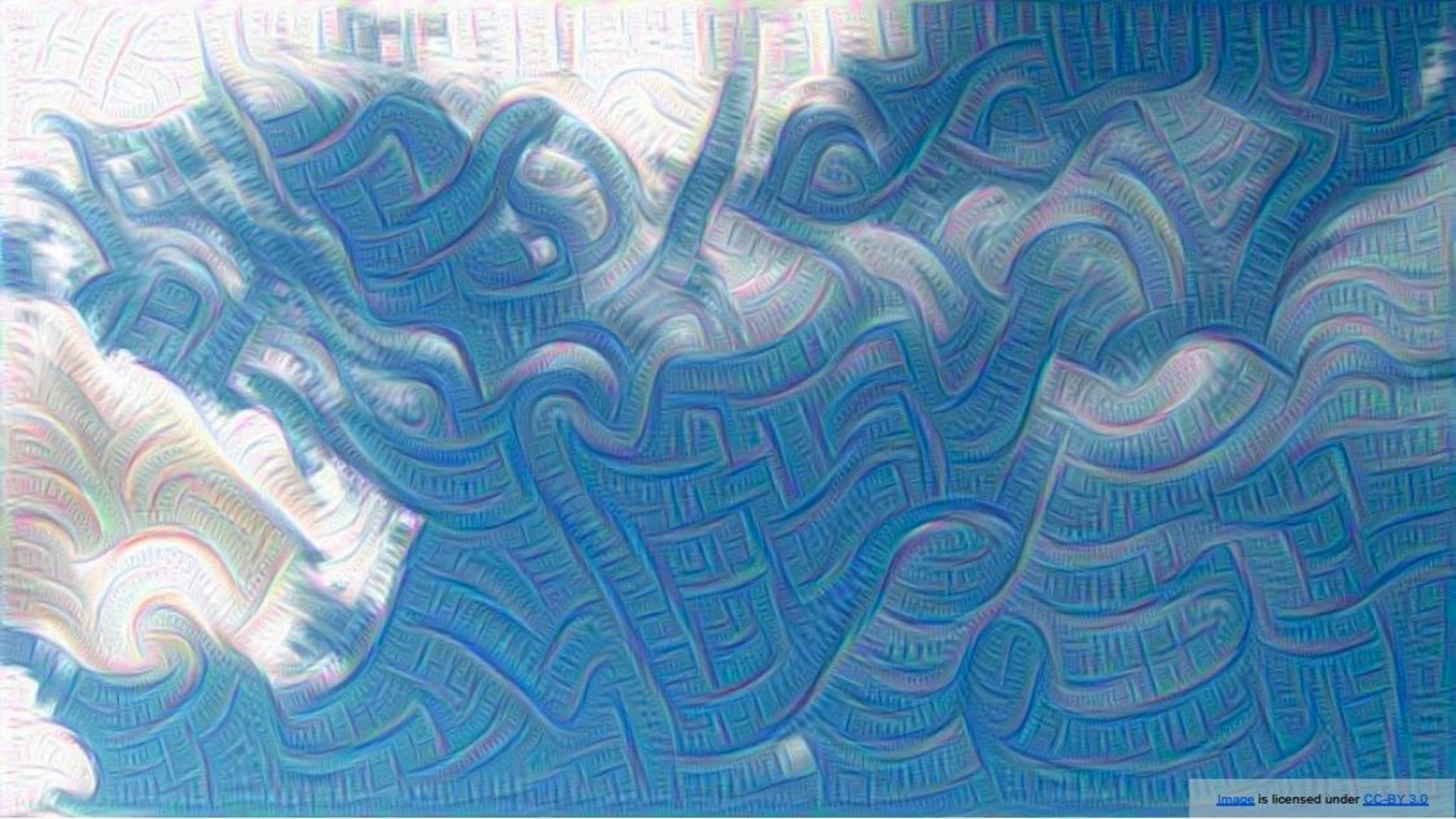
Google DeepDream

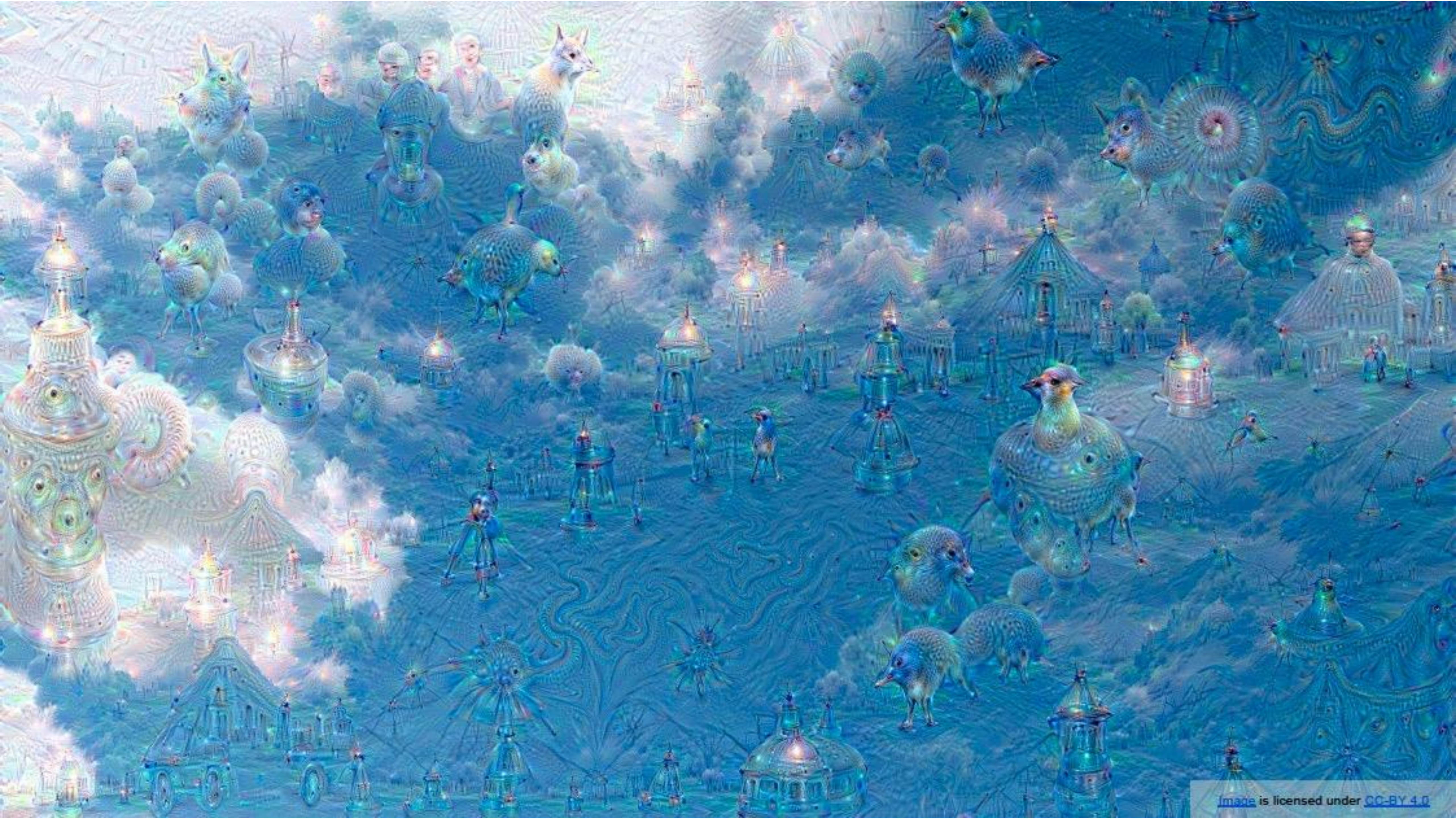
Amplify one **layer** instead of just one neuron.

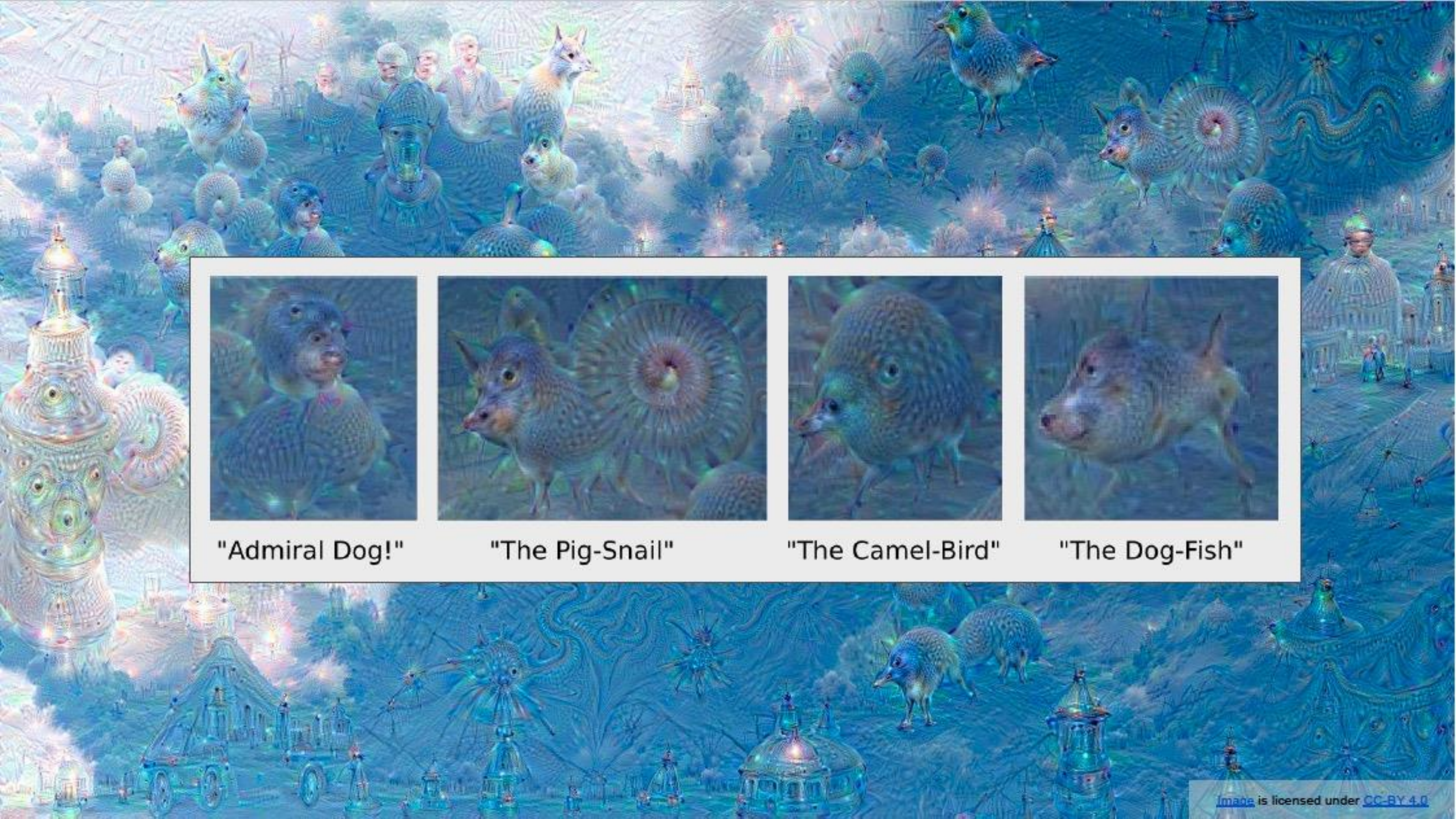


Choose an image and a layer in a CNN; repeat:

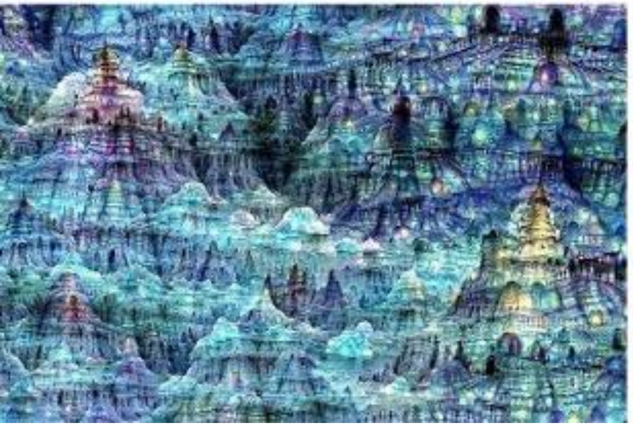
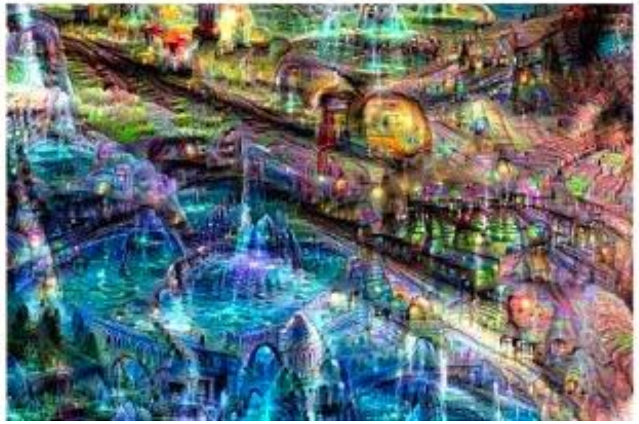
1. Forward: compute activations at chosen layer
2. Set gradient of chosen layer *equal to its activation*
   Equivalent to maximizing $\sum_i f_i^2(x)$
3. Backward: Compute gradient w.r.t. image
4. Update image (with some tricks)





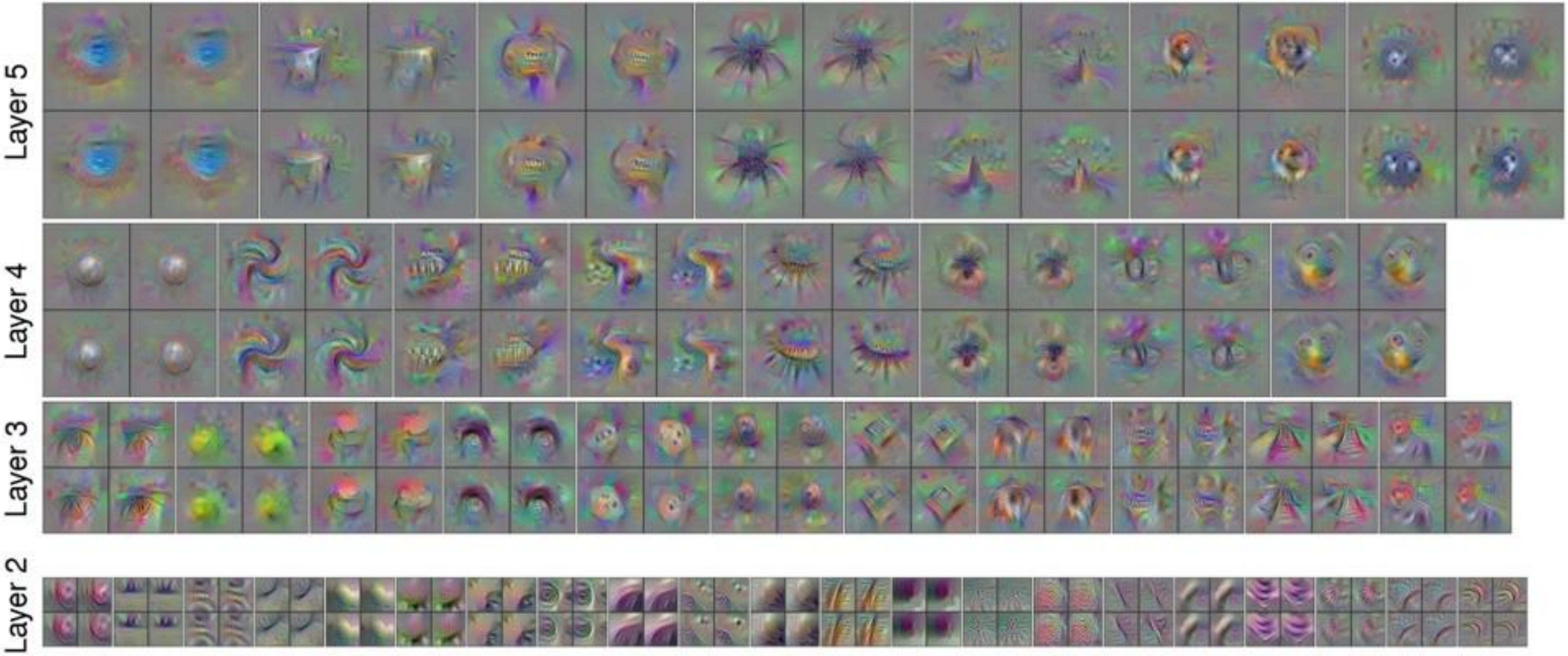

"Admiral Dog!"
"The Pig-Snail"
"The Camel-Bird"
"The Dog-Fish"



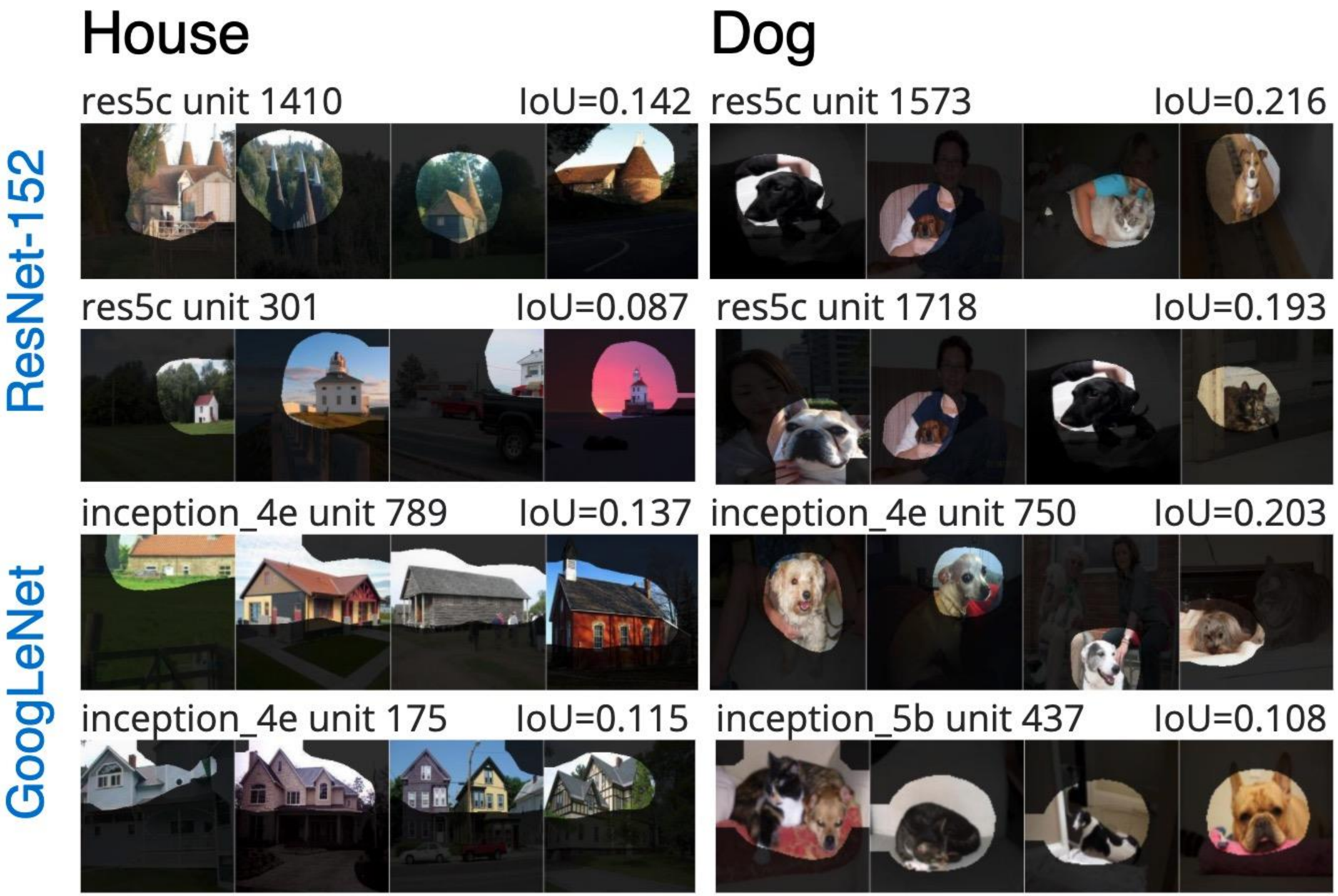



Quantification on the units

Quantifying interpretability of units

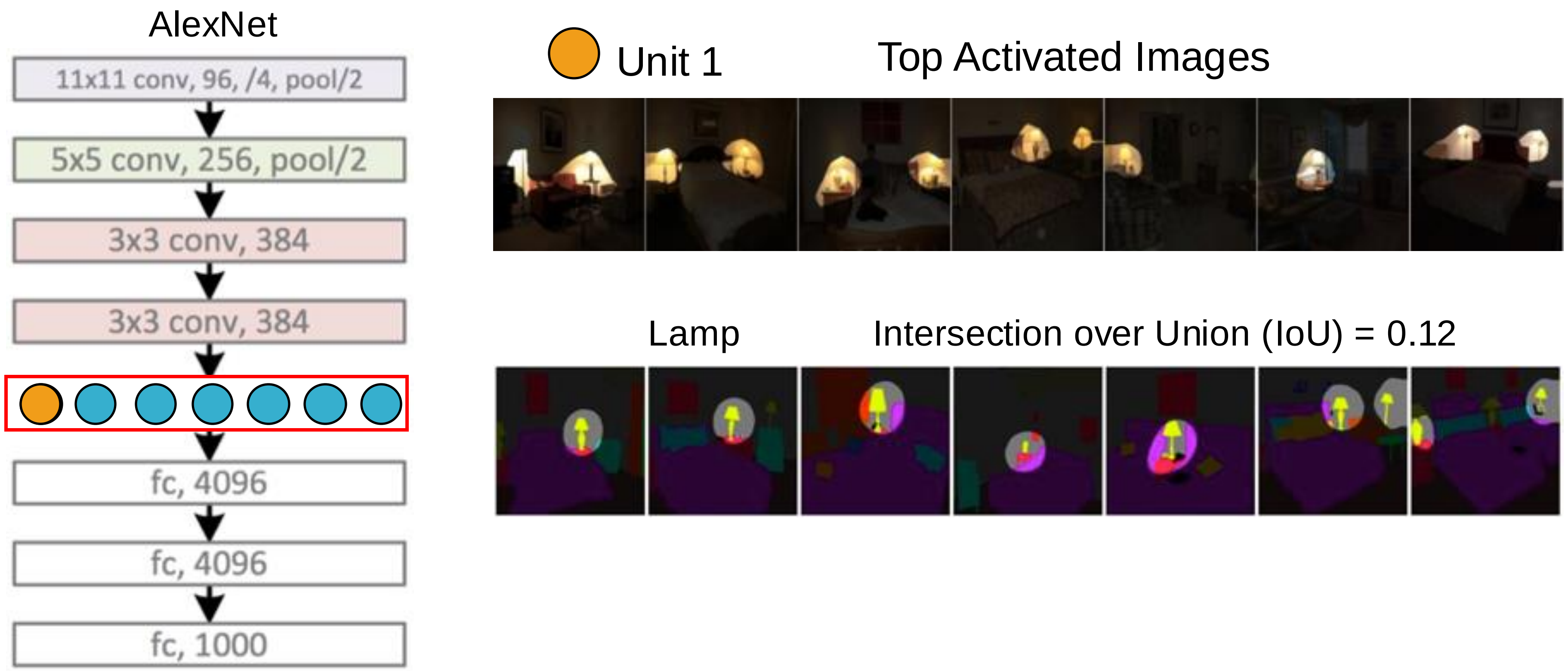


Visualization



Quantifying interpretability of units

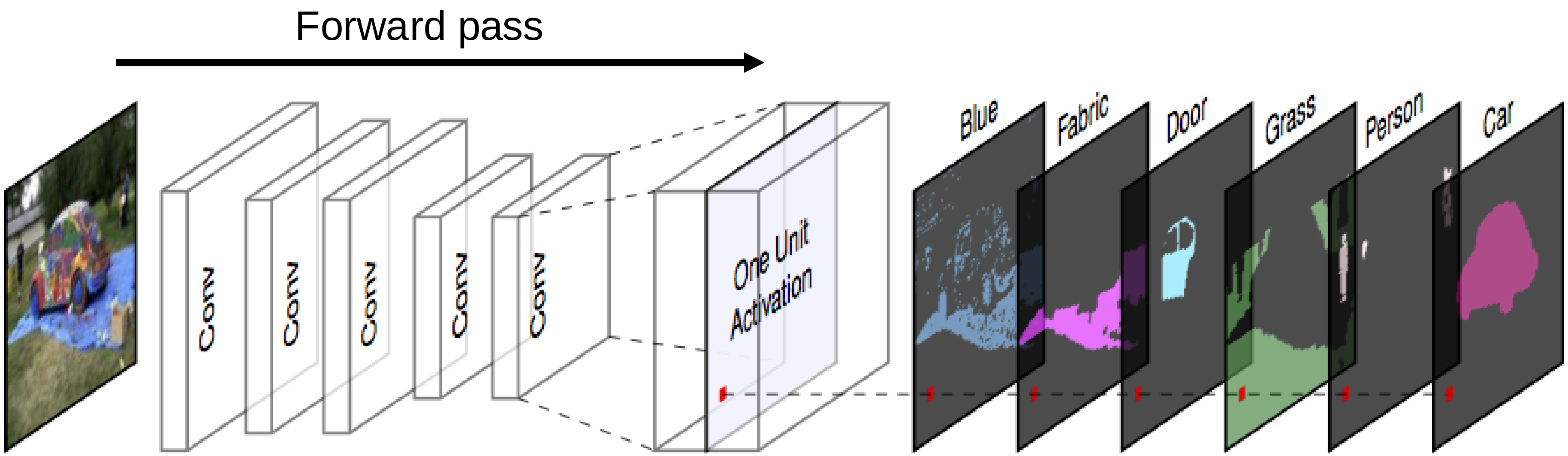
- For a given unit, measure the overlap between areas of high activation and semantic segmentations for a large set of visual concepts



Bau et al., 2017

Quantifying interpretability of units

- For a given unit, measure the overlap between areas of high activation and semantic segmentations for a large set of visual concepts

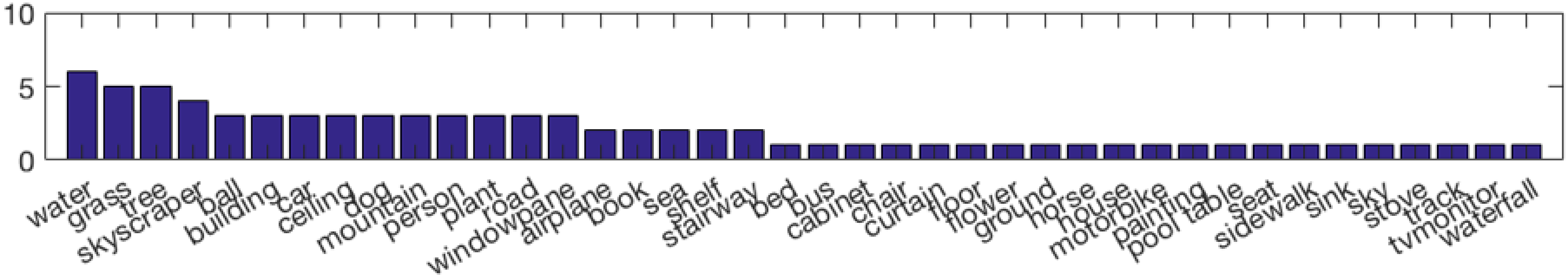


Network trained for image classification on ImageNet or Places

Dataset with semantic segmentations for ~1200 visual concepts

Quantifying interpretability of units

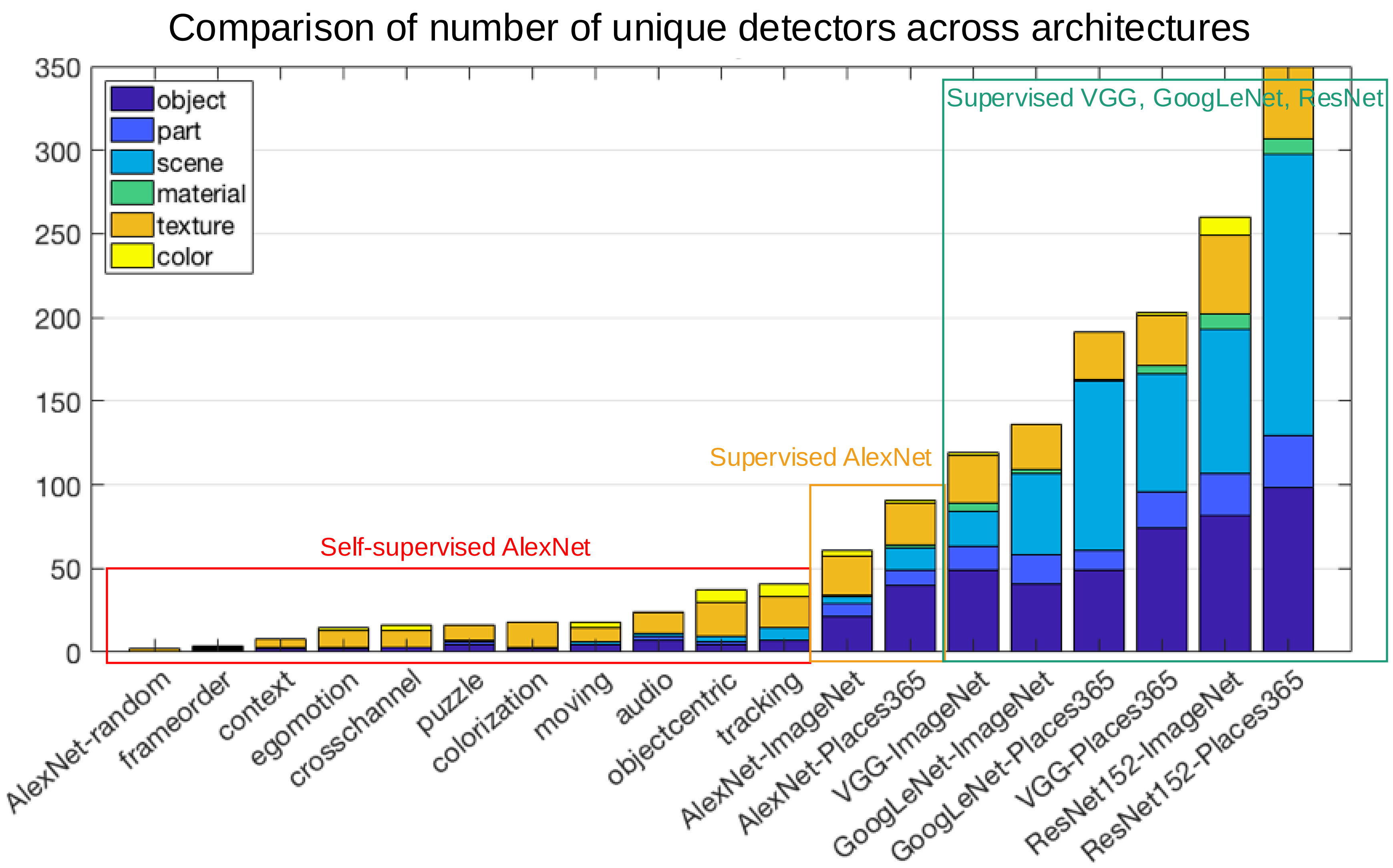
Histogram of object detectors for Places AlexNet conv5 units
81/256 units with IoU > 0.04



conv5 unit 79 car (object) IoU=0.13

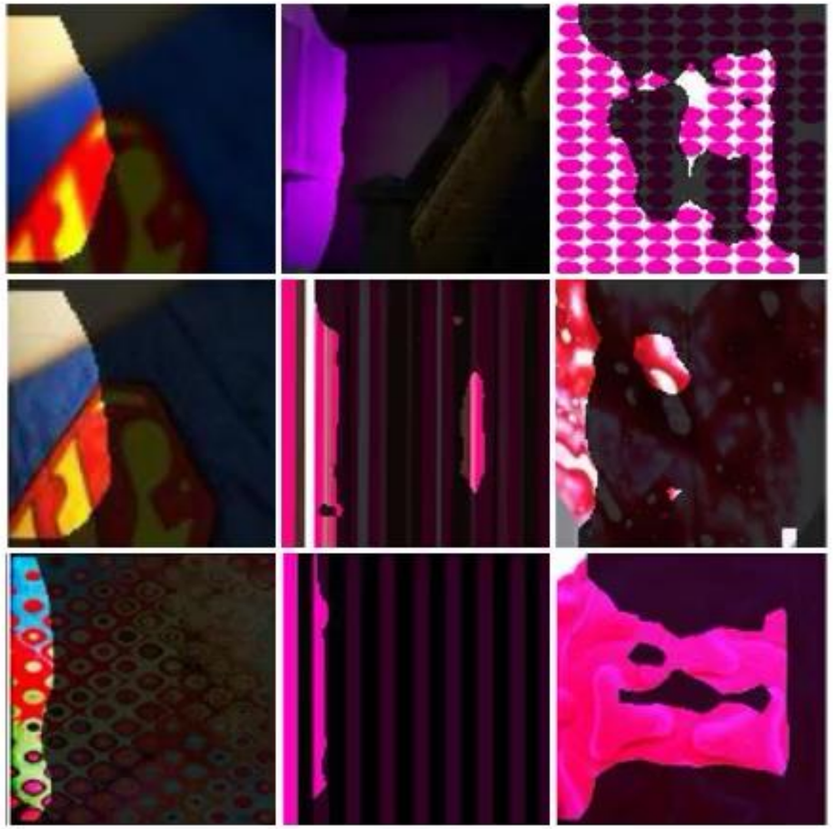
conv5 unit 107 road (object) IoU=0.15

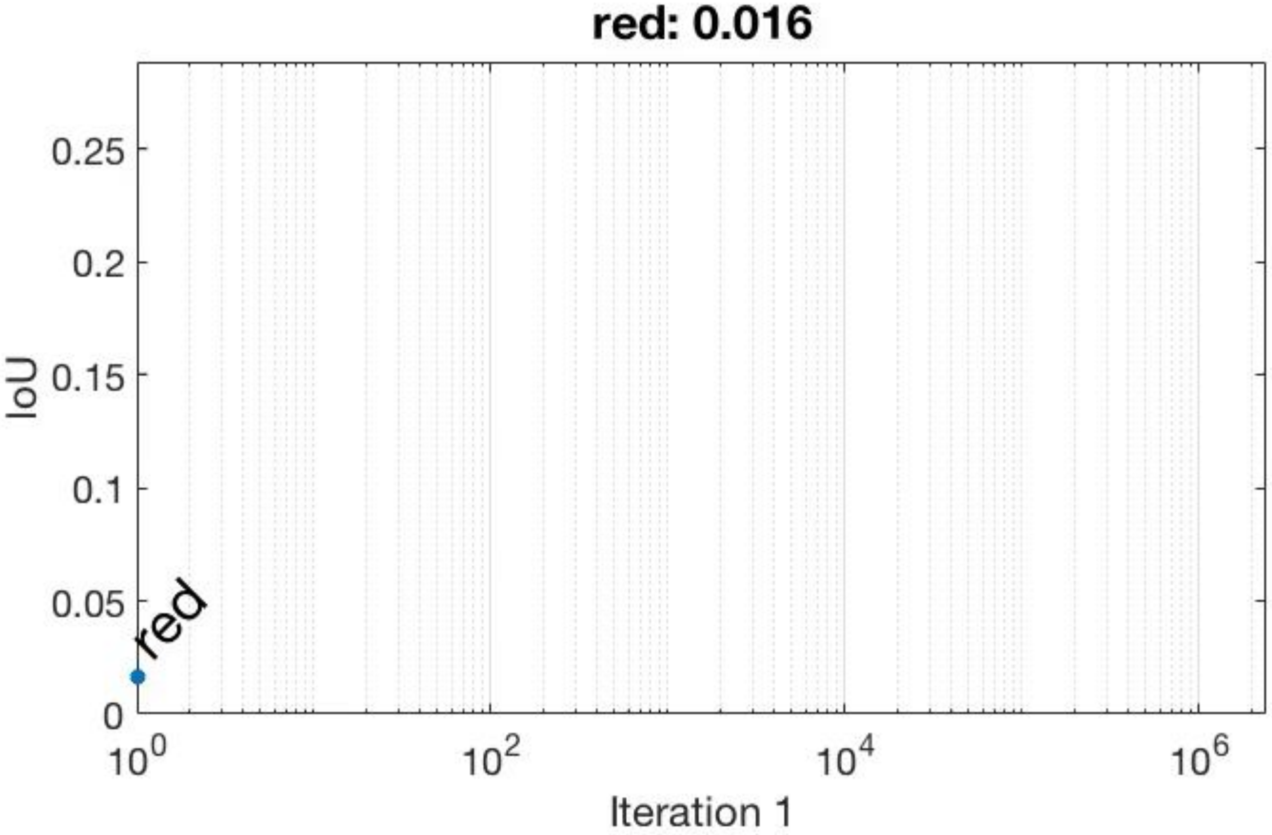
Quantifying interpretability of units



Change of Unit during Finetuning

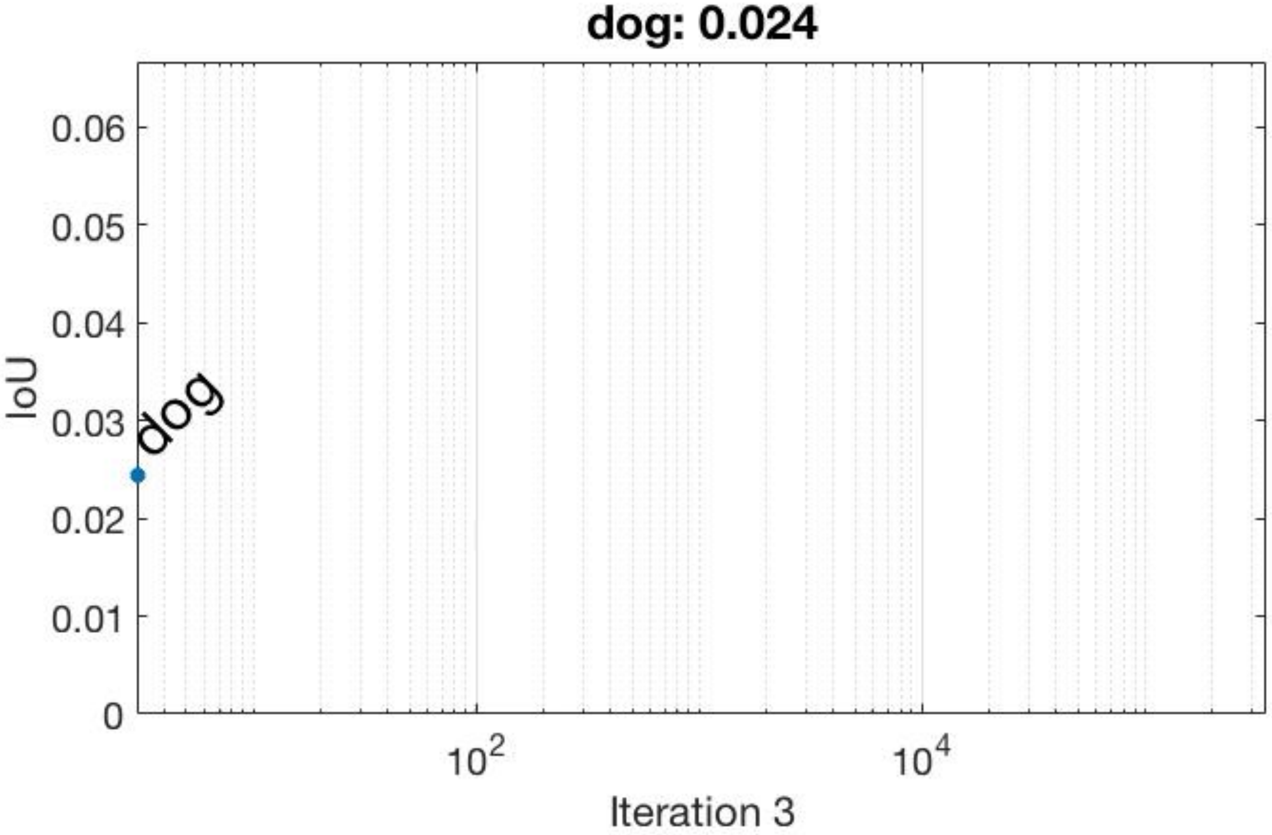
When learning from scratch, units change from detectors for low-level patterns or visually simple concepts such as "road" to detectors for more complex higher-level concepts such as "car".





Change of Unit during Finetuning

When fine-tuning a model to solve a new problems, internal units can change roles. This "dog" detector becomse a "waterfall" detector when retrained from object to scene classification.



This Class

- Basics
- Visualizing “Saliency map”
- Visualization by maximizing activation
- Quantification on the units

Next Class

Object Detection