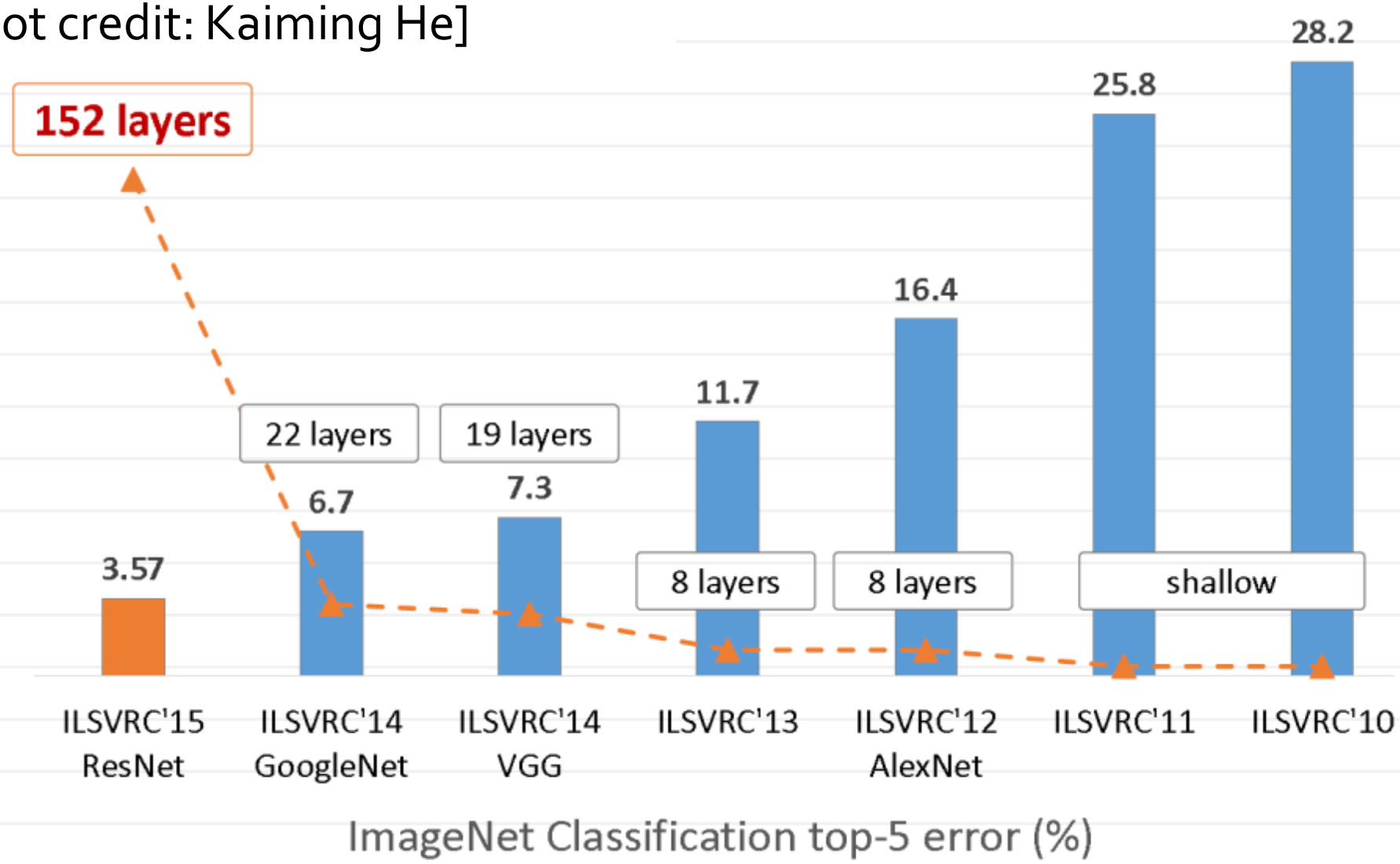
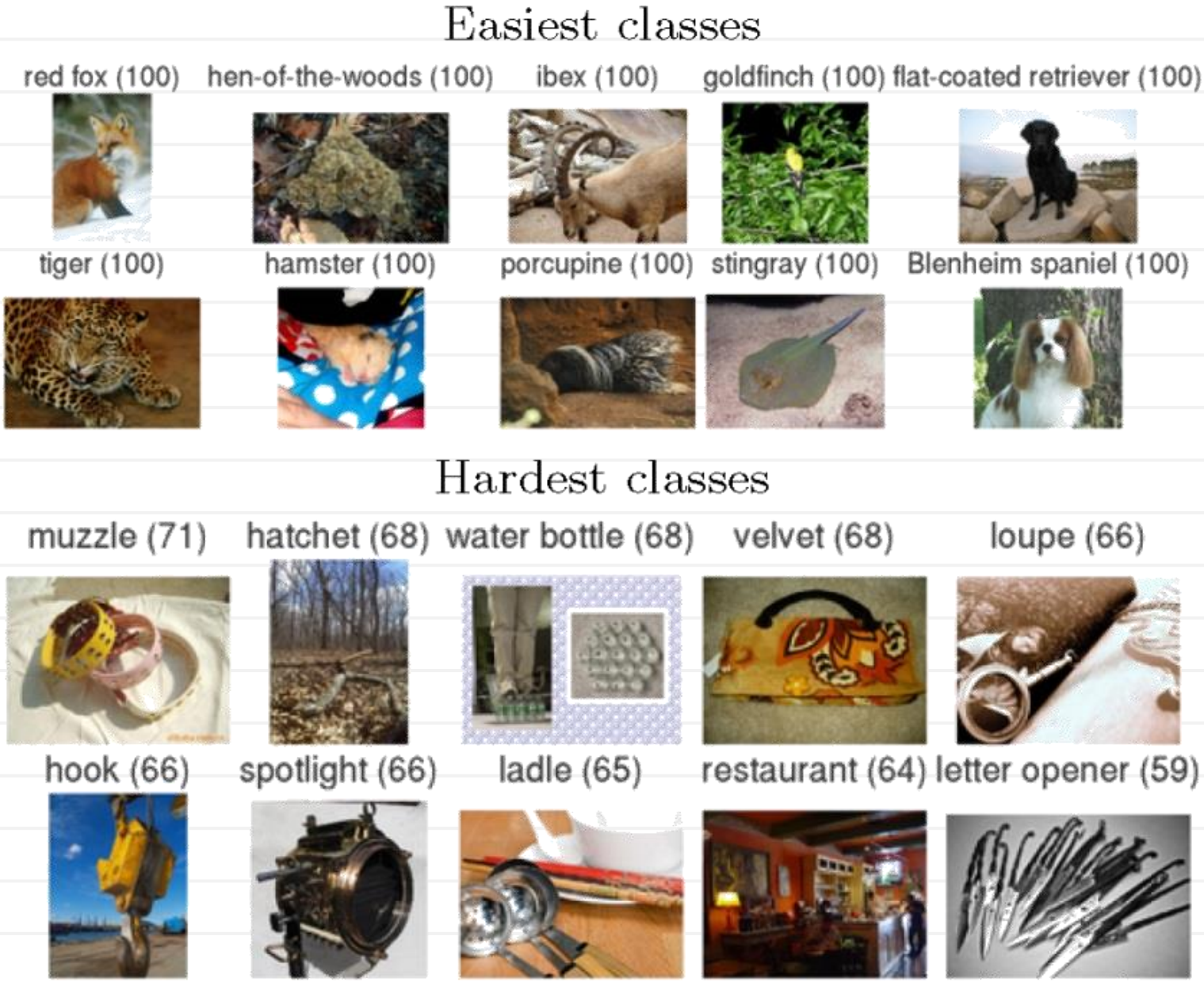
Lecture 4: Representations inside ConvNets

Building the best network

[plot credit: Kaiming He]



Easy and hard cases



[Russakovsky et al. 2014]

Texture beats shape as a cue

(a) Texture image

81.4%	**Indian elephant**
10.3%	indri
8.2%	black swan

(b) Content image

71.1%	**tabby cat**
17.3%	grey fox
3.3%	Siamese cat

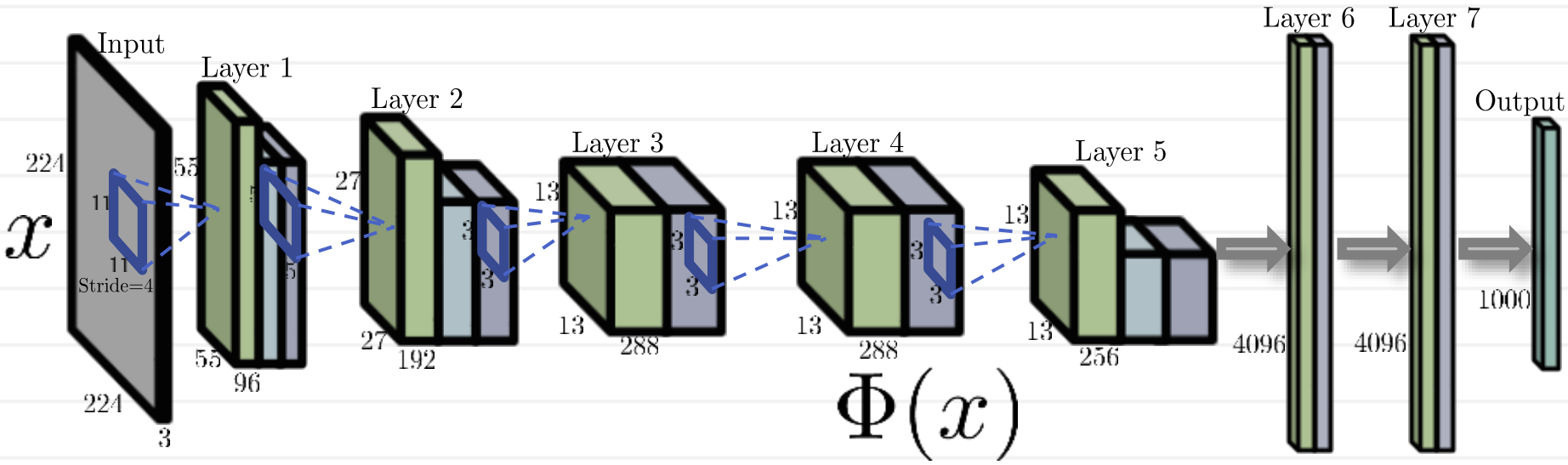
(c) Texture-shape cue conflict

63.9%	**Indian elephant**
26.4%	indri
9.6%	black swan

(this is indri)

[Geirhos et al. ICLR19]

Representations inside the neural network

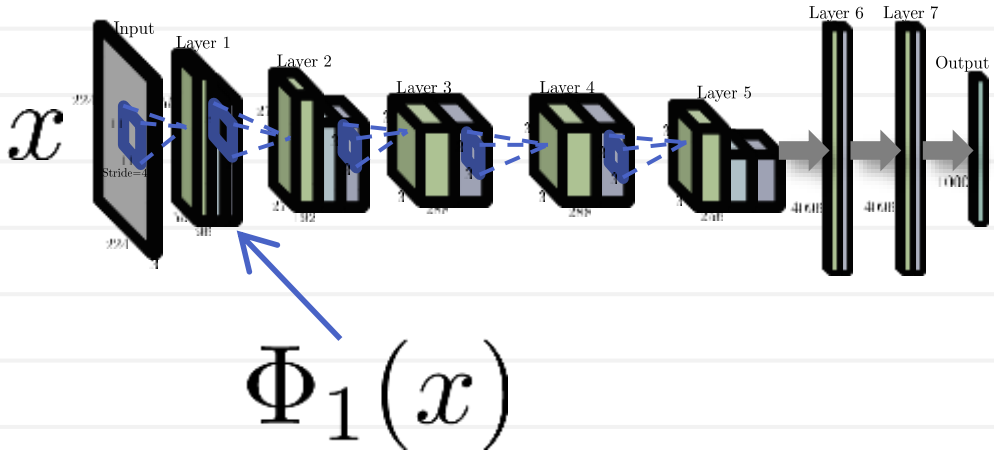


Lots of important questions:

- What are their properties?
- Are they redundant?
- Are they invertible?
- **Are they useful?**

Pattern sensitivity

Types of patterns in each layer:



$$\arg\max_{x \in S} \max_{p,q} \Phi_1(x)^t [p,q]$$

[Zeiler Fergus 14]

Pattern sensitivity

Types of patterns in each layer:

Layer 1

Layer 2

[Zeiler Fergus 14]

Pattern sensitivity

Types of patterns in each layer:

[Zeiler Fergus 14]

Layer 3

Pattern sensitivity

Types of patterns in each layer:

Layer 4

Layer 5

[Zeiler Fergus 14]

Meaningful units or meaningful space?

Max activations for random units:



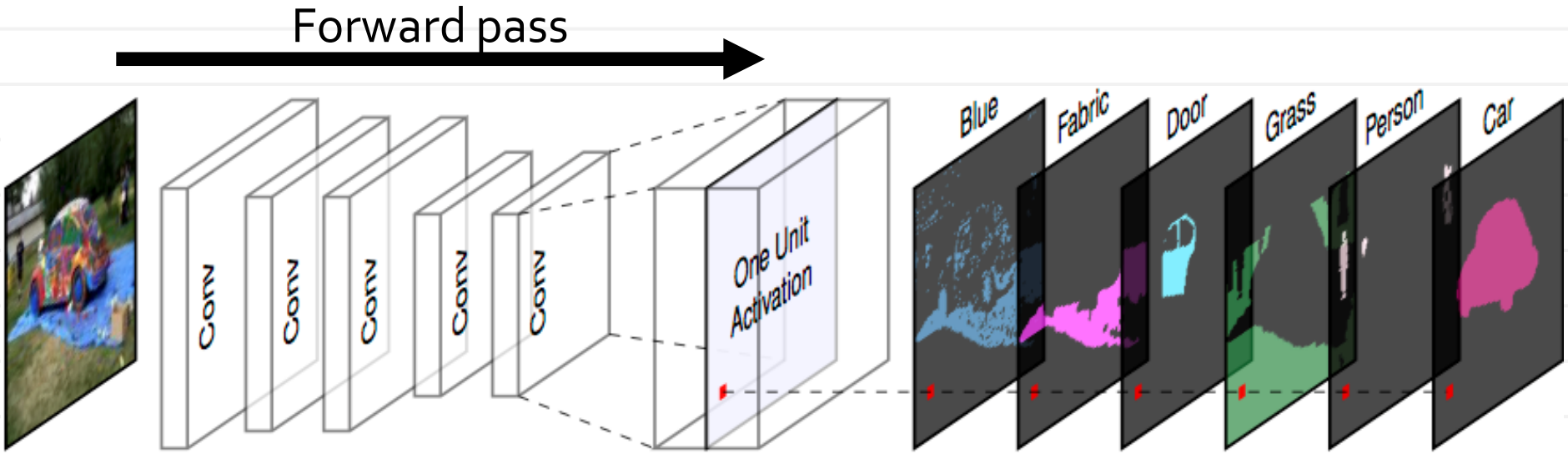

Max activations for random directions:

$$\arg\max_{x \in S} \sum_t w_t \Phi_4(x)^t [p, q]$$




[Szegedy et al. 2014]

Network dissection [Zhou et al. CVPR17]



Total = 63,305 images
1,197 visual concepts

ADE20K Zhou et al, CVPR'17
Pascal Context Mottaghi et al, CVPR'14
Pascal Part Chen et al, CVPR'14
Open-Surfaces Bell et al, SIGGRAPH'14
Describable Textures Cimpoi et al, CVPR'14
Colors

Network dissection [Zhou et al. CVPR17]

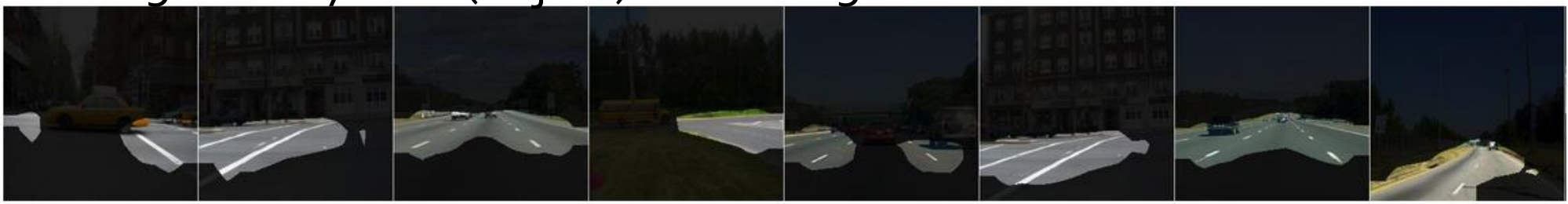
activation mask (top 5%)

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

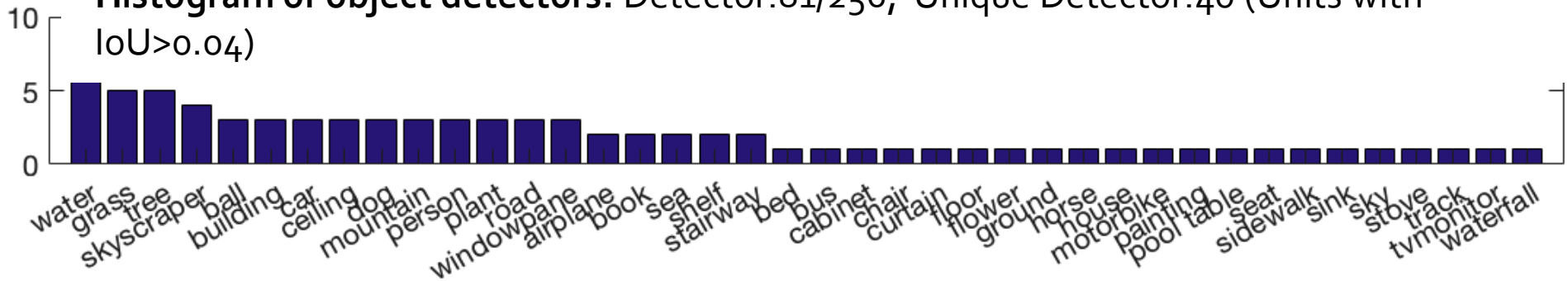
concept mask

conv5 unit 79 car (object) IoU=0.13

conv5 unit 107 road (object) IoU=0.15

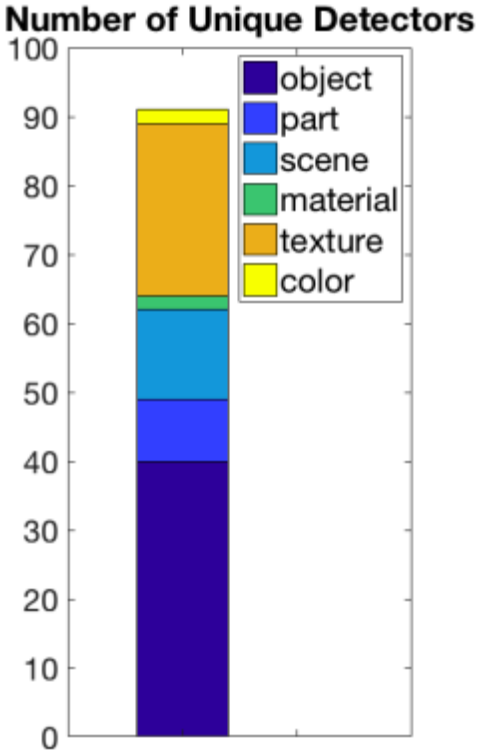


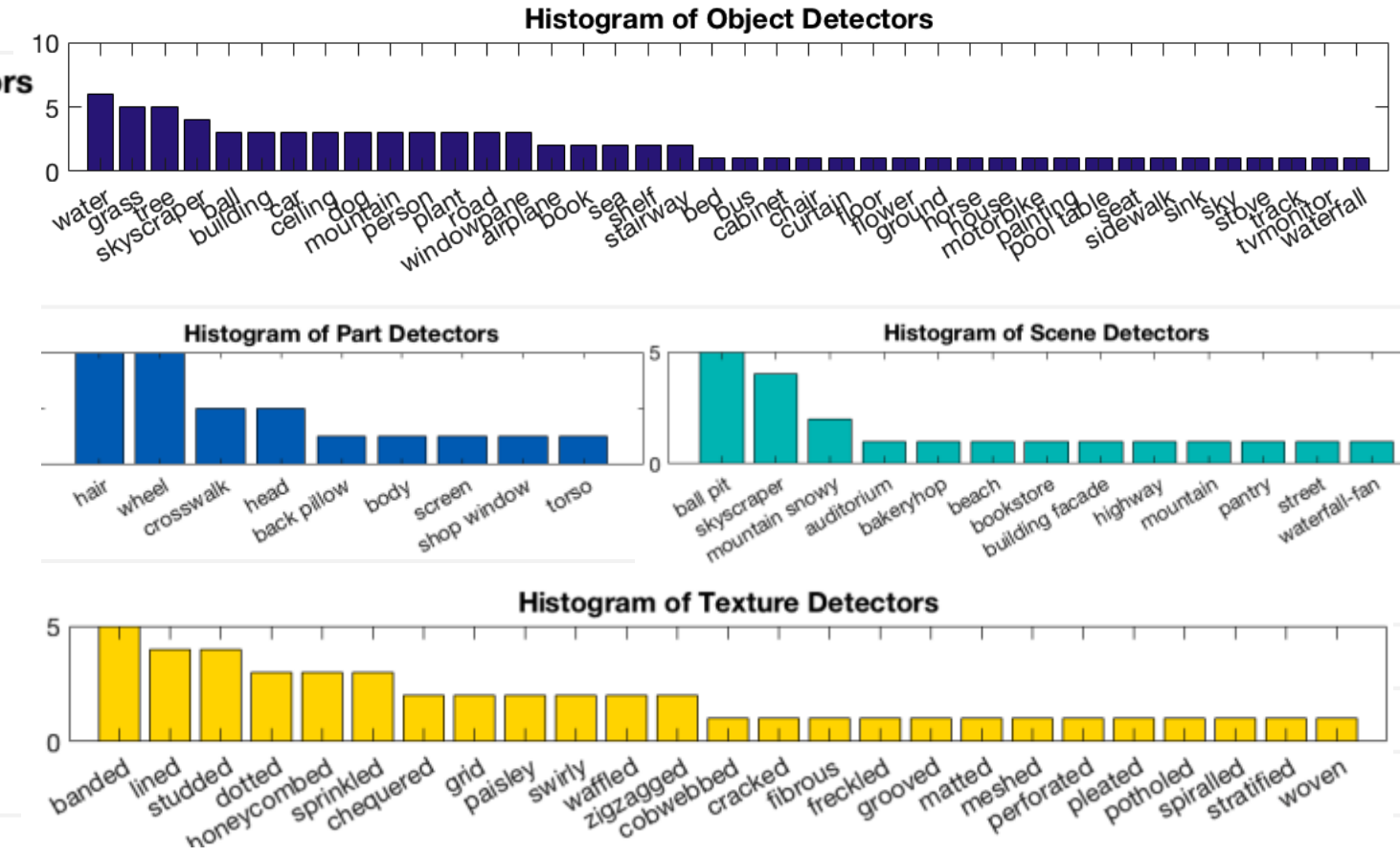
Histogram of object detectors: Detector:81/256, Unique Detector:40 (Units with IoU>0.04)



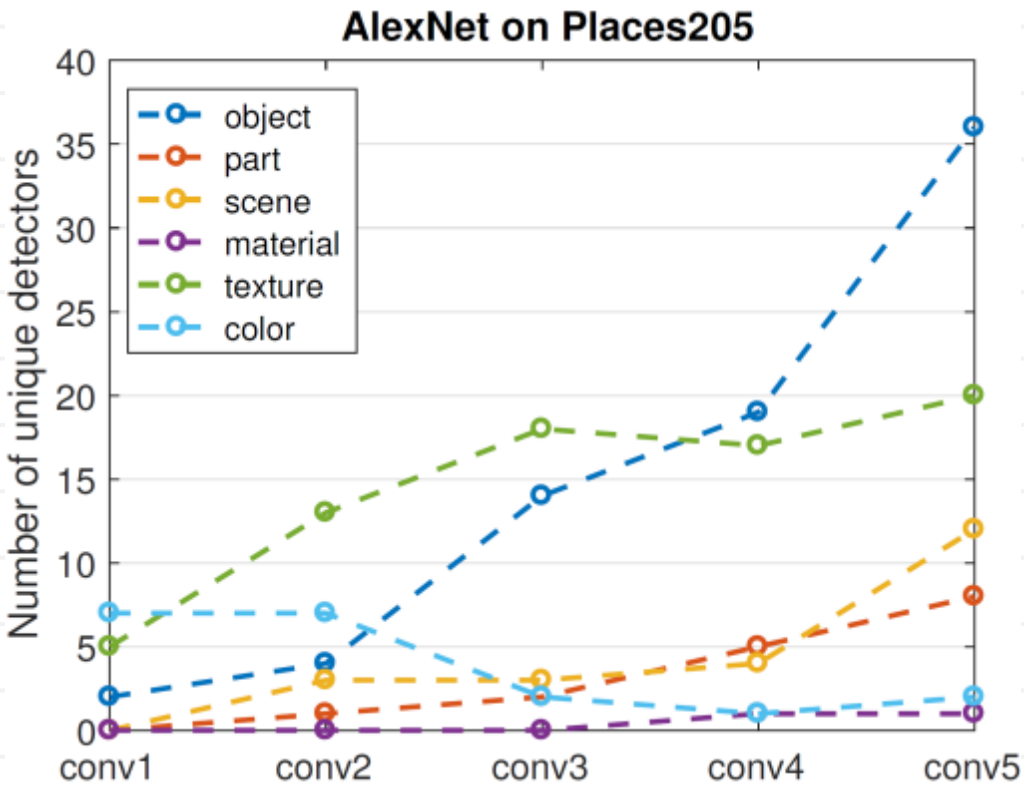
Network dissection [Zhou et al. CVPR17]

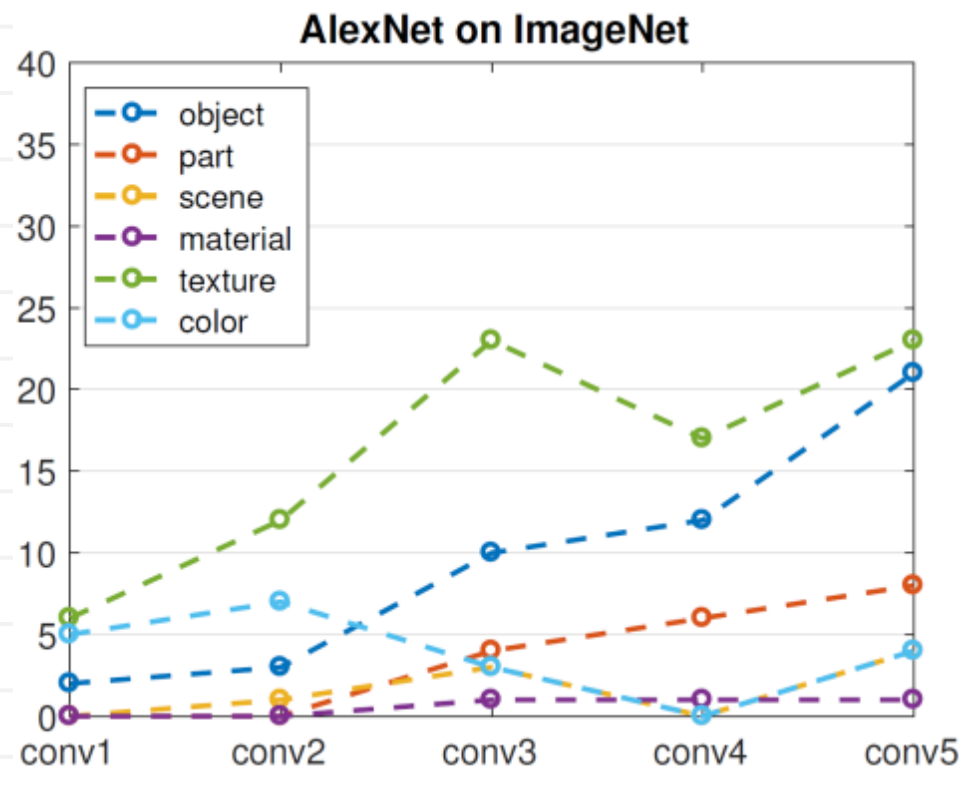
Dissection Report





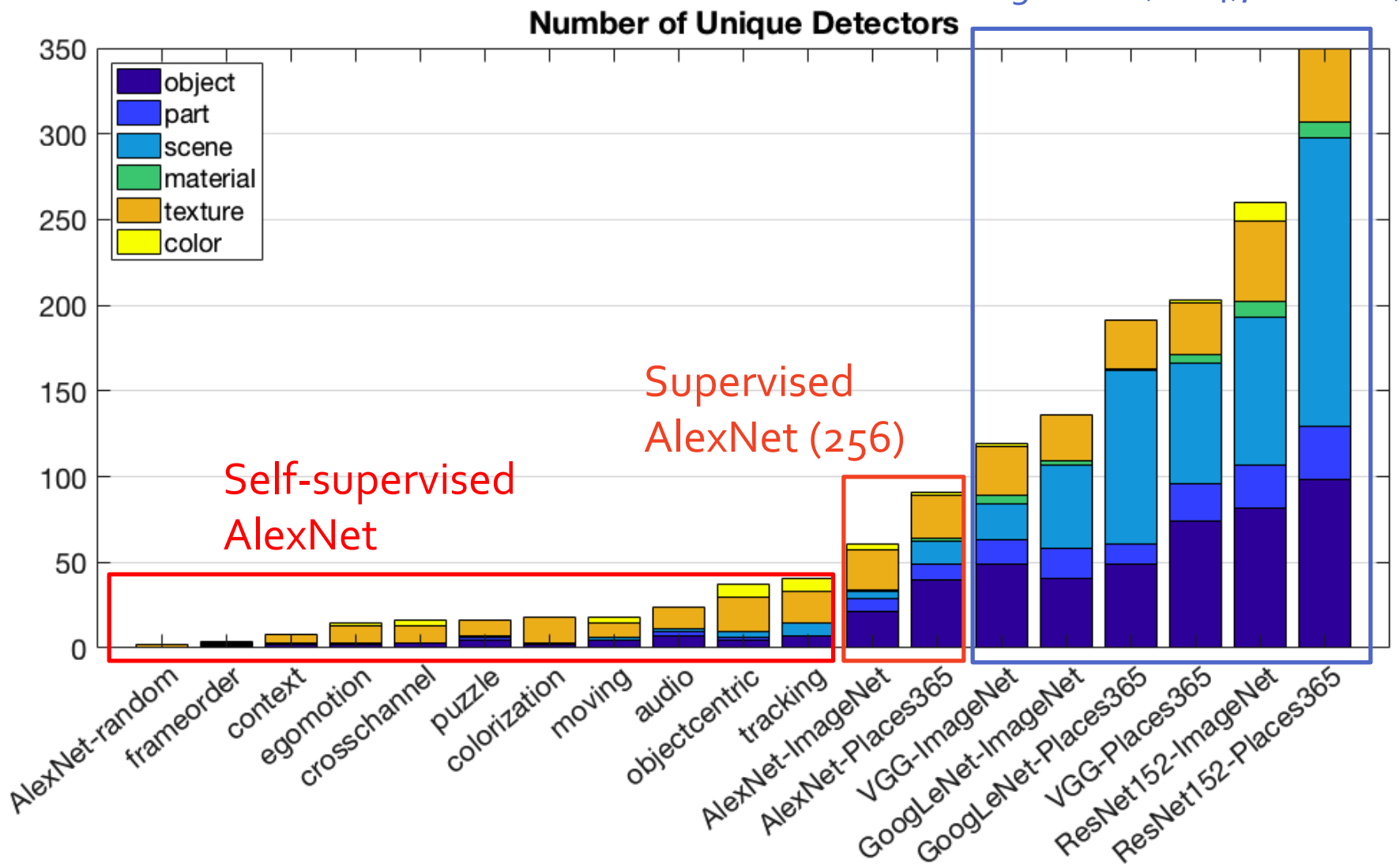
Network dissection [Zhou et al. CVPR17]



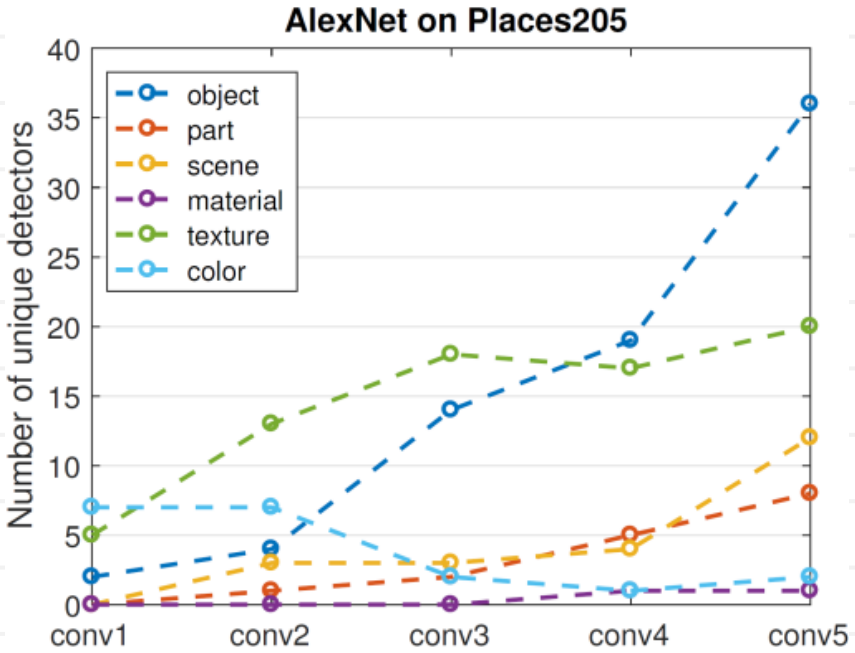


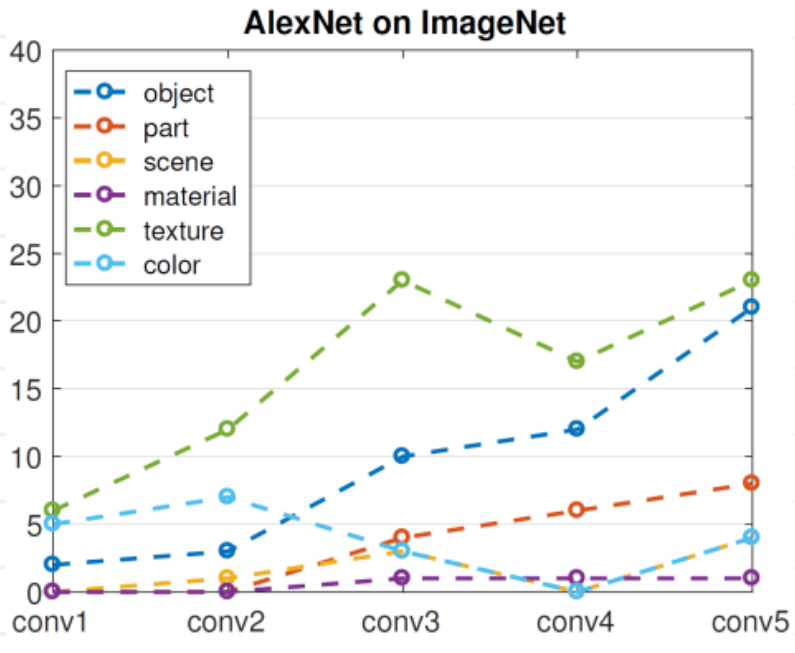
Network dissection [Zhou et al. CVPR17]

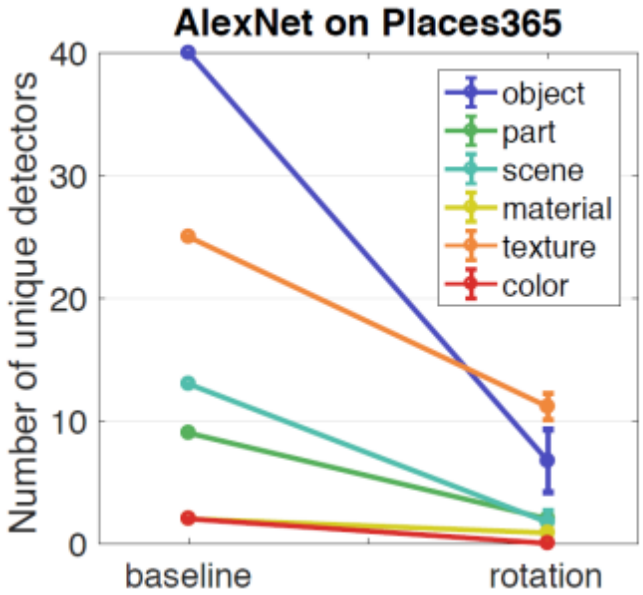
Comparing last conv layers:

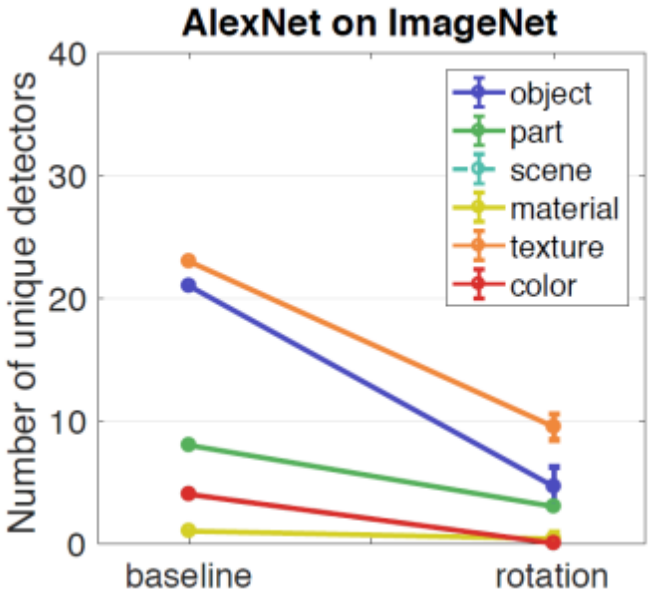


Network dissection [Zhou et al. CVPR17]

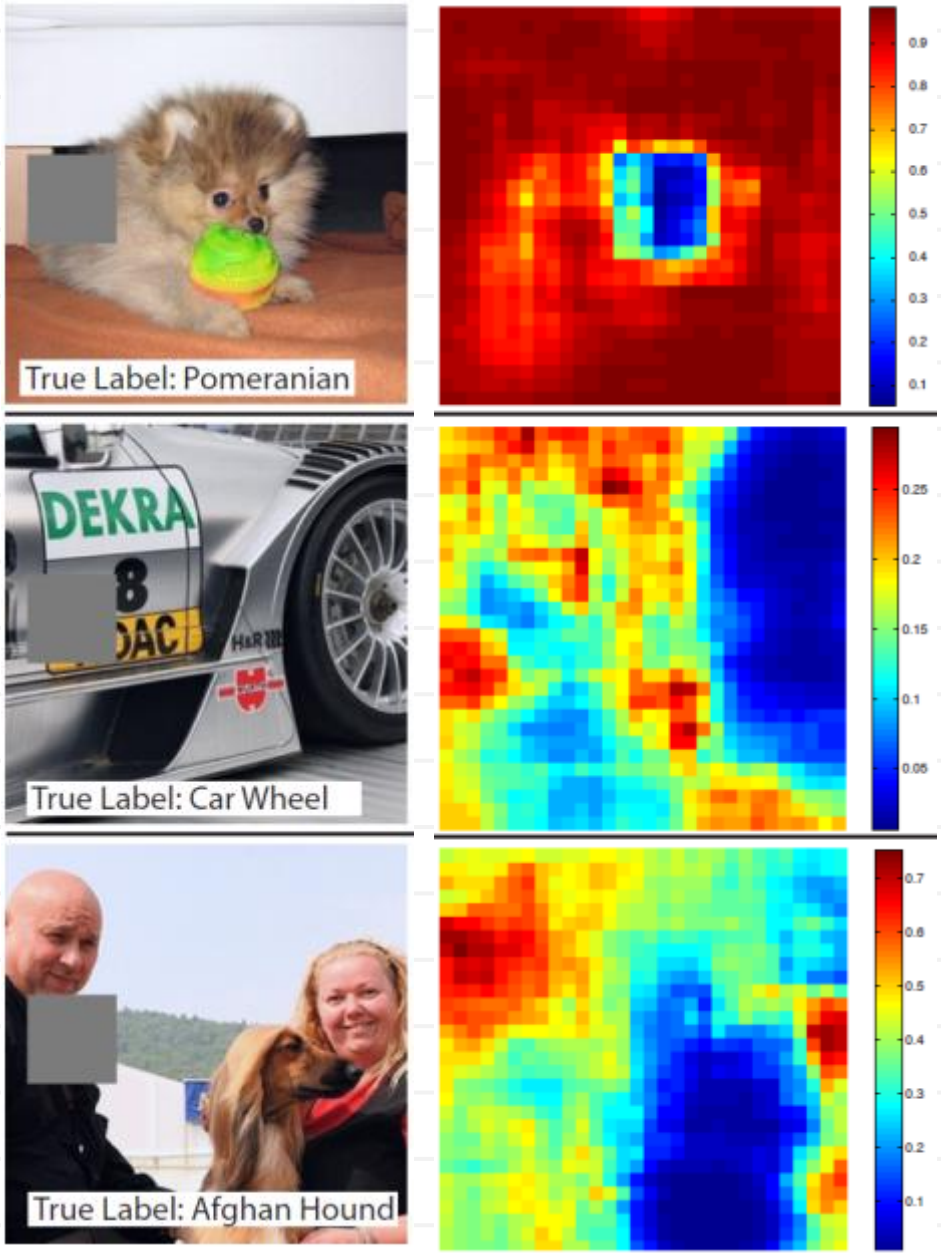








Grounding CNN decisions

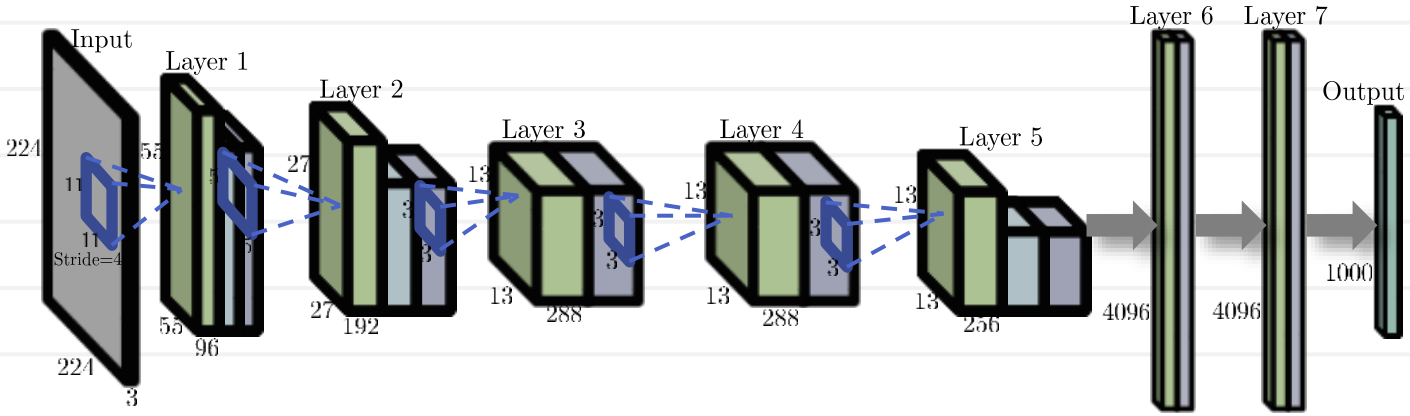


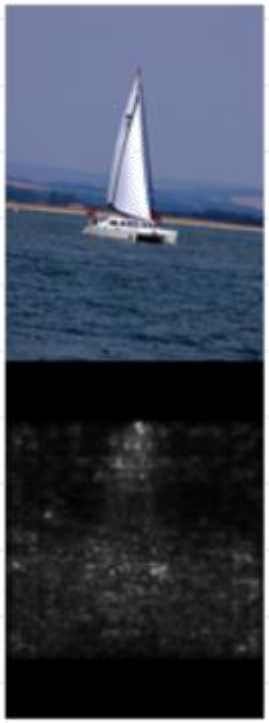
1. Occluding different regions
2. Visualizing the score
3. Good visualizations, but slow
4. Arbitrary choice of occluder

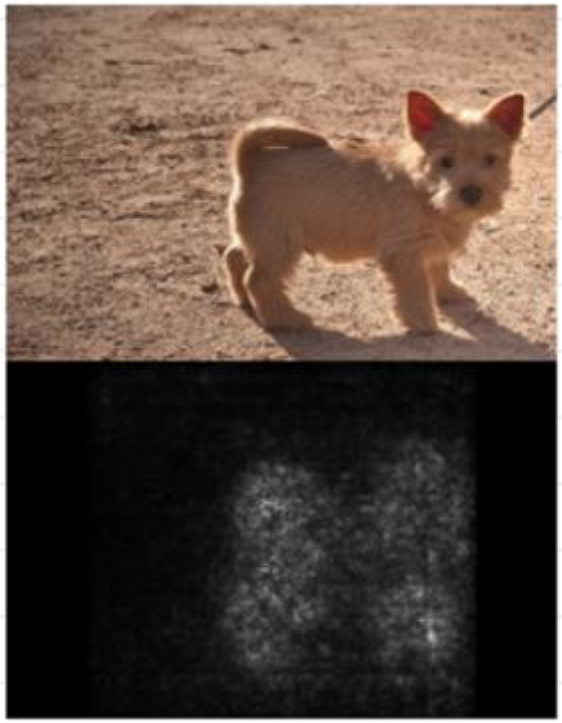
Can we do grounding faster?

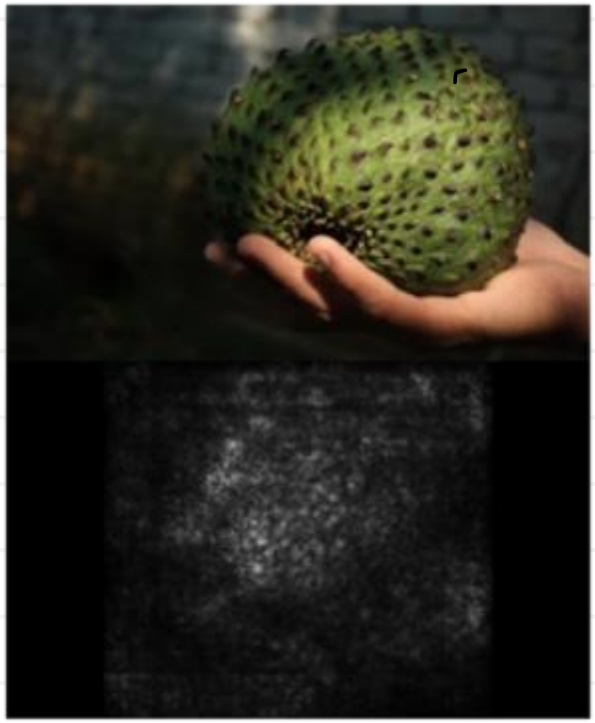
Grounding using gradient

$$\left\| \frac{\partial \Phi_{\text{Last}}[y_0]}{\partial x} \right\|$$



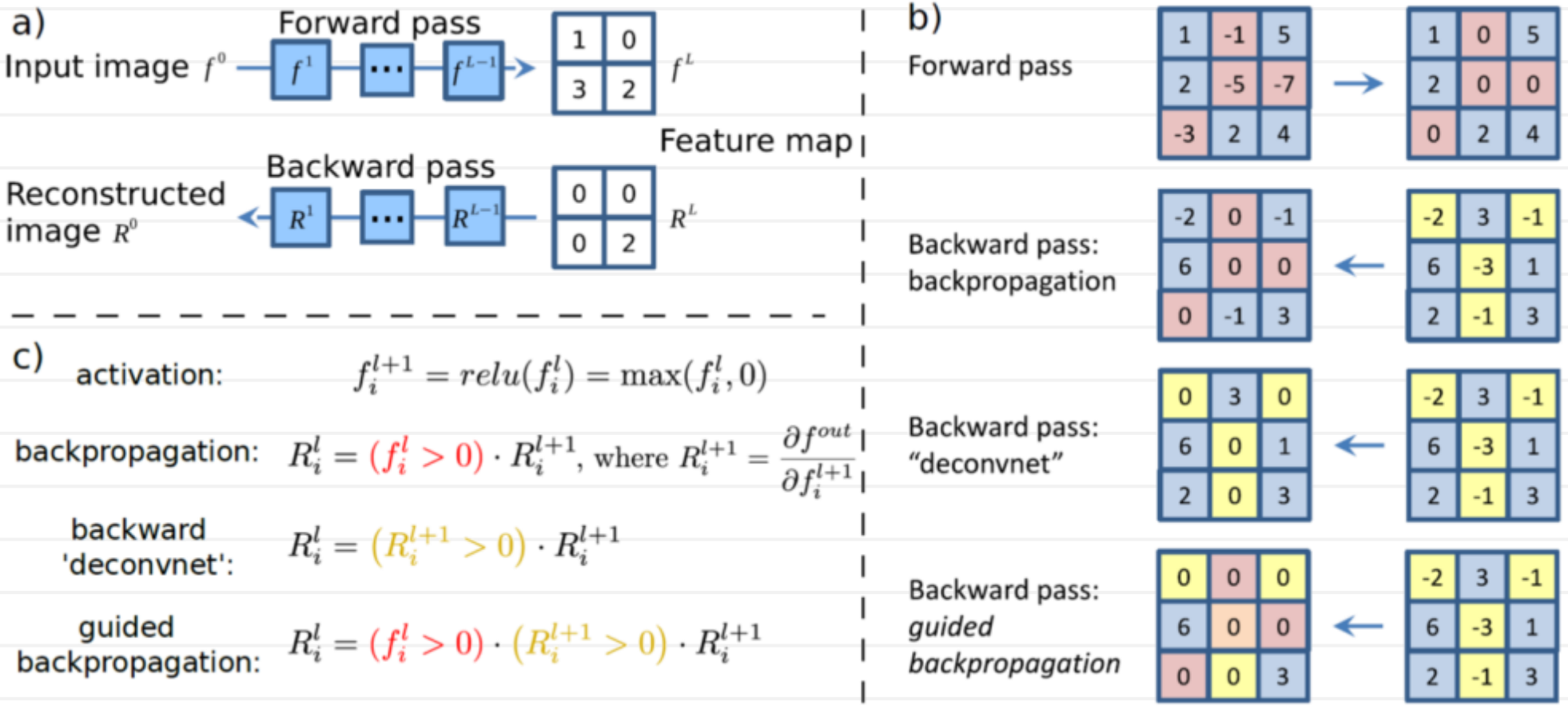






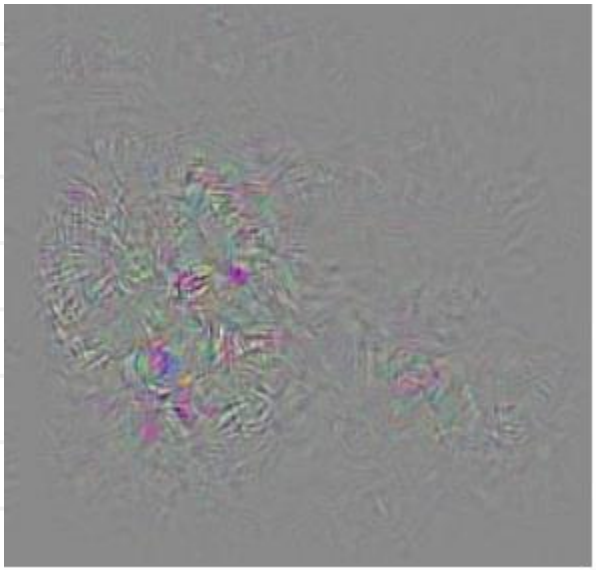
[Simonyan et al. 2013]

Better grounding using gradient

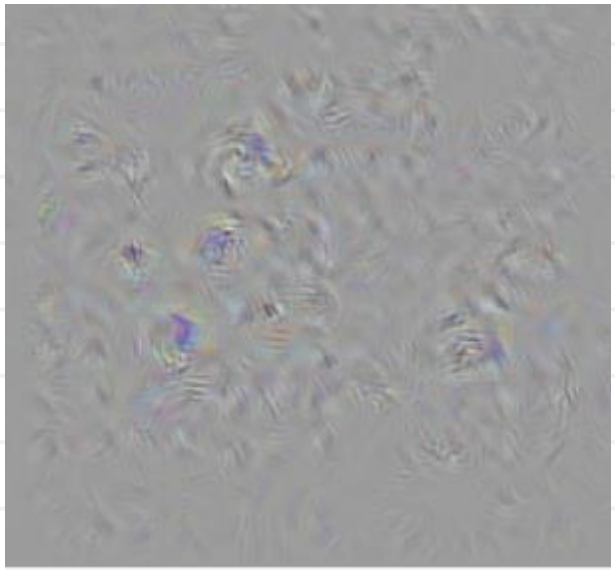


[Springenberg et al. 2015]

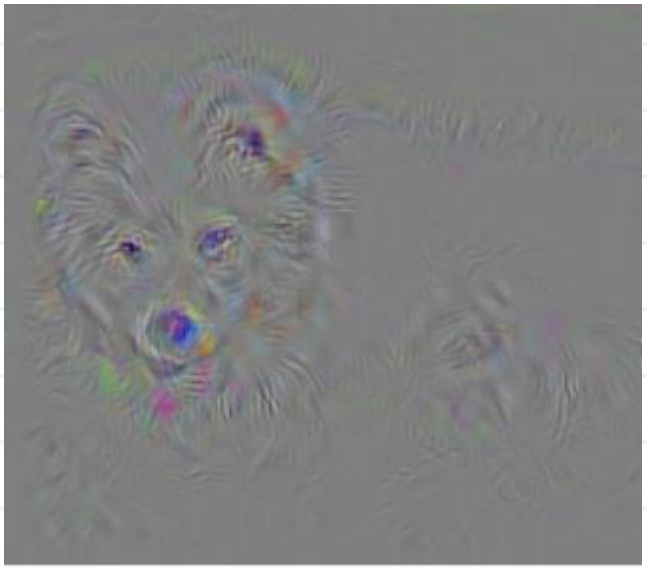
Better grounding using gradient



gradient



DeConvNet

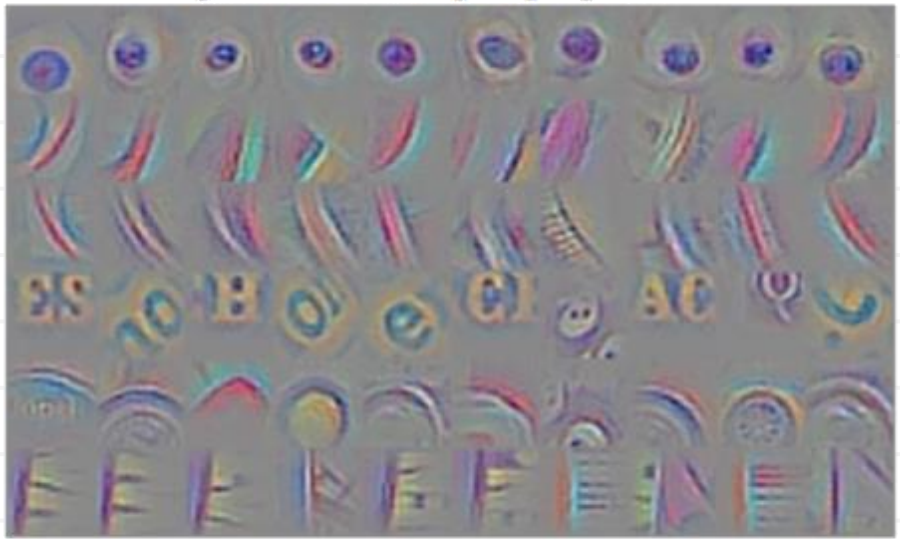


Guided backprop

[Springenberg et al. 2015]

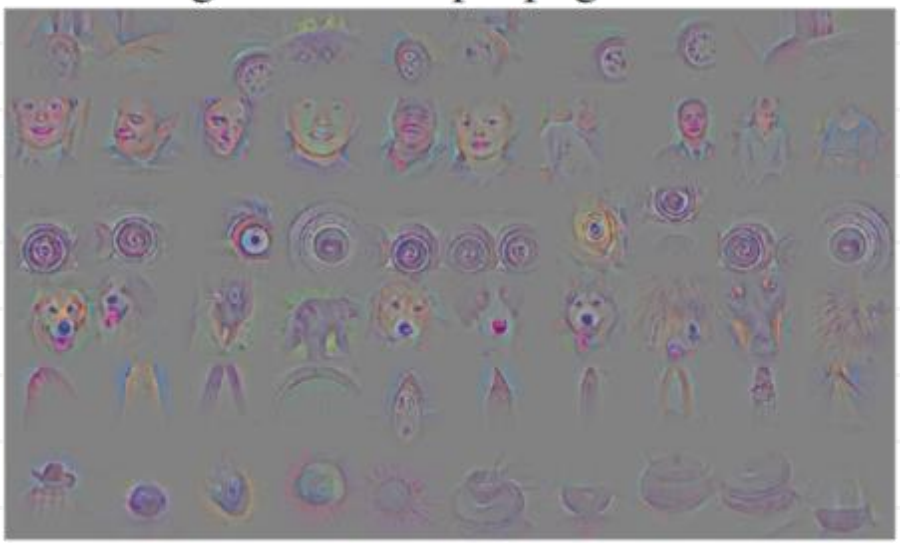
Better grounding using gradient

guided backpropagation

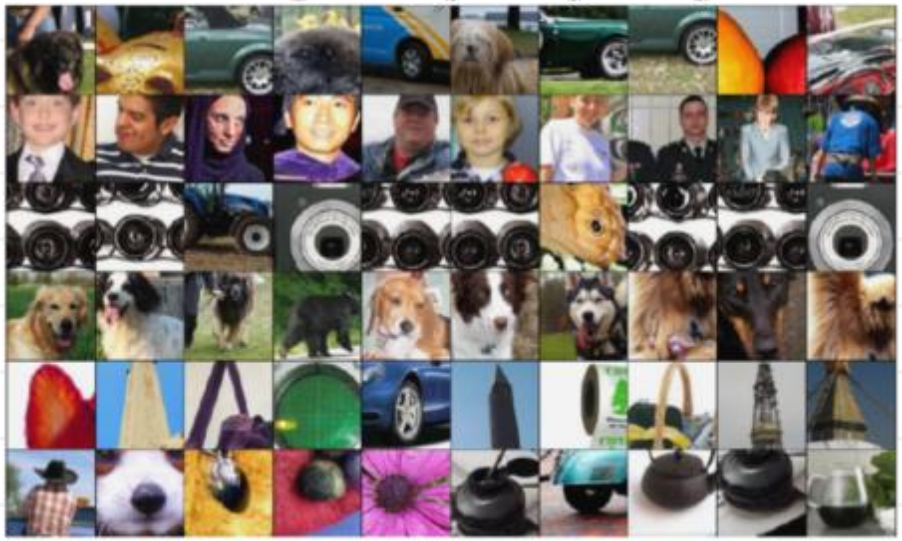


corresponding image crops

guided backpropagation



corresponding image crops



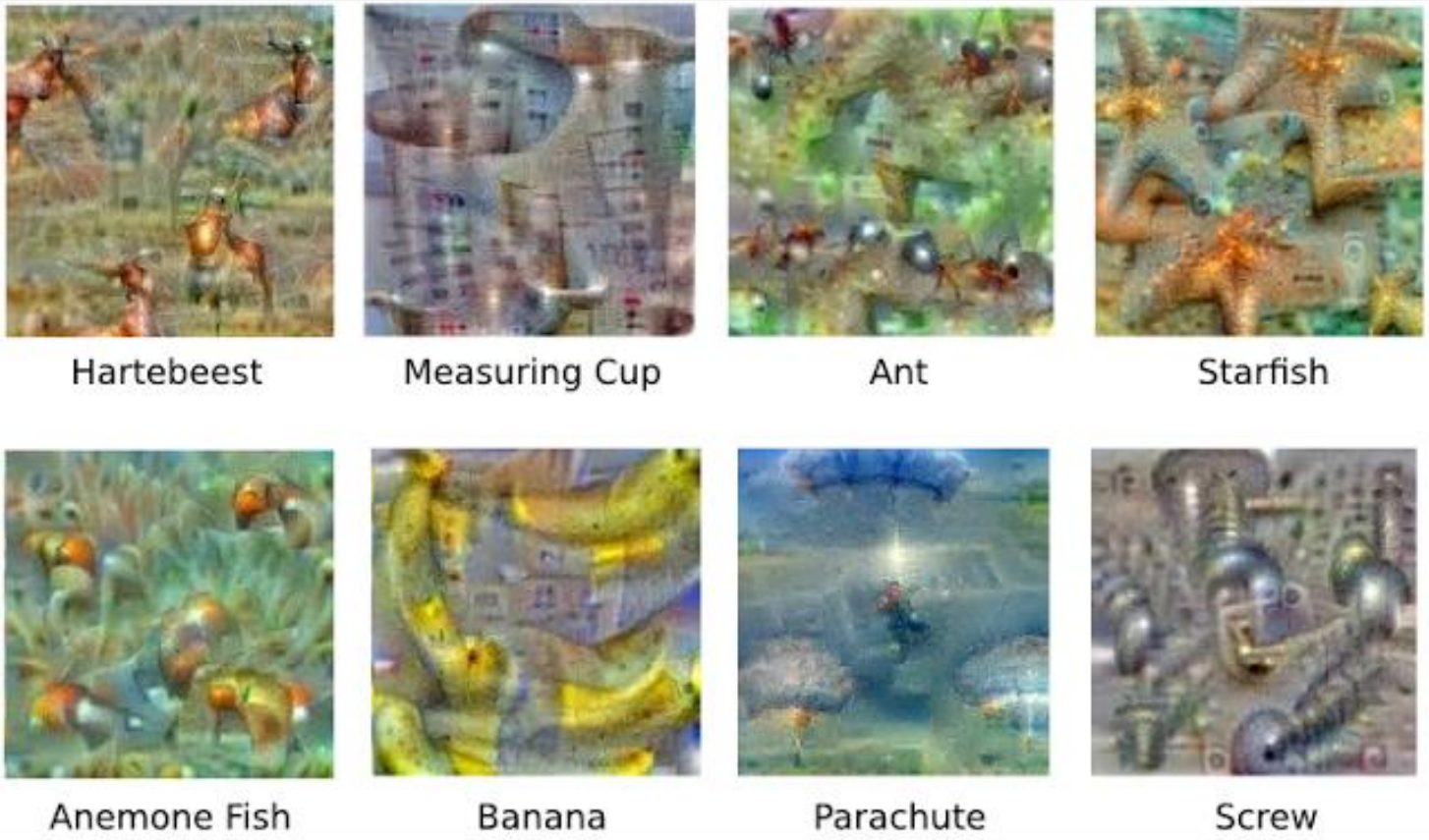
[Springenberg et al. 2015]

Maximum impulses

$$\hat{x} = \arg\max_{x} \left(\Phi_{\text{Last}}(x)[y_0] - \lambda R(x) \right)$$

[Simonyan et al. 2013]

Inceptionism: maximum impulses



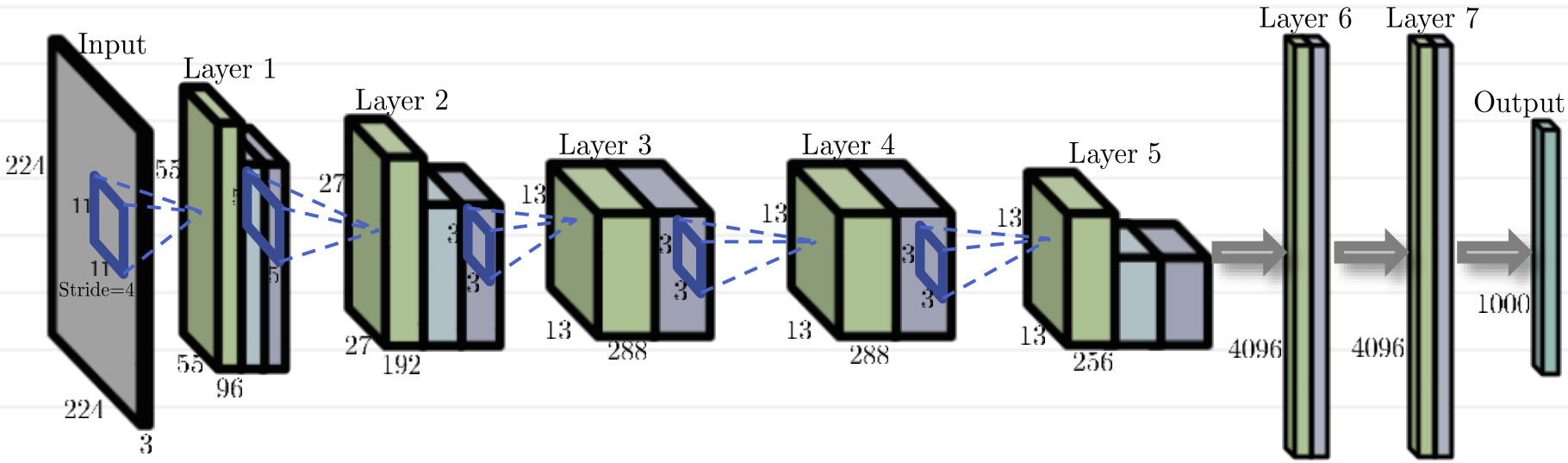
Deeper network + "Smart" regularizer (jitter) + maximizing logits (pre-softmax)

https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

Inceptionism: maximum impulses



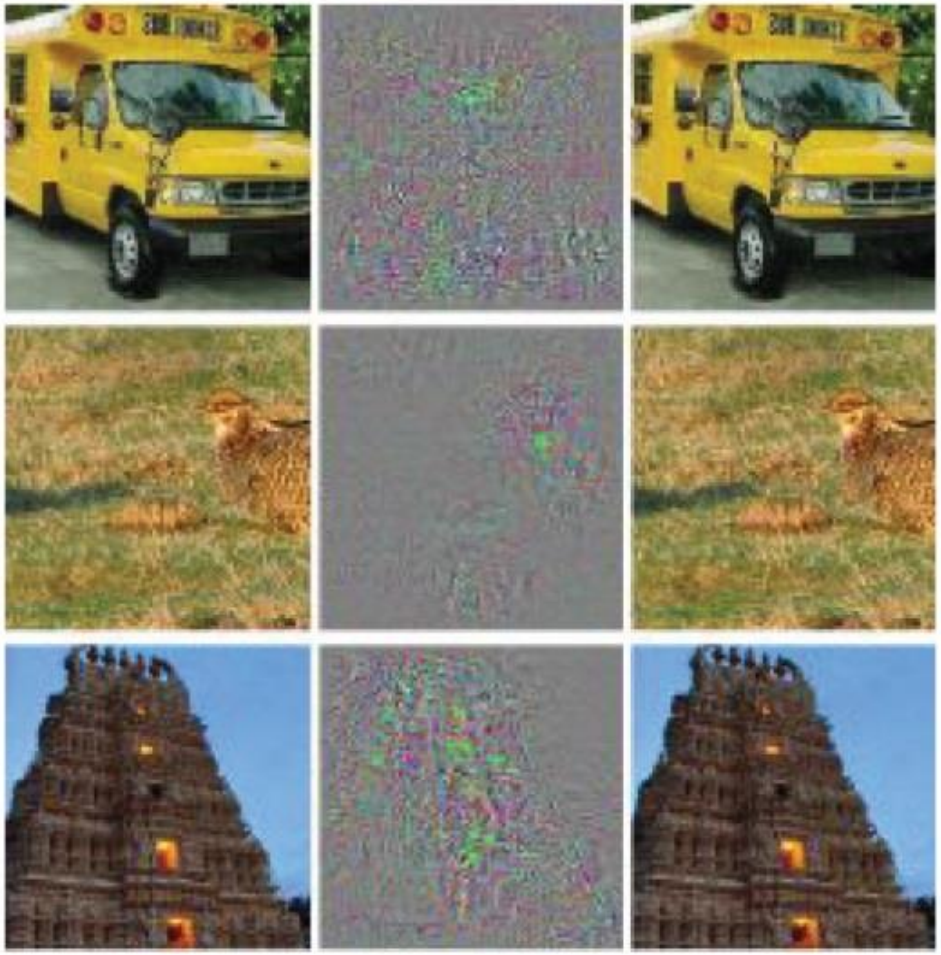

Generating adversarial perturbations

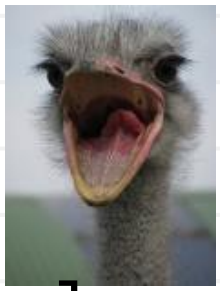


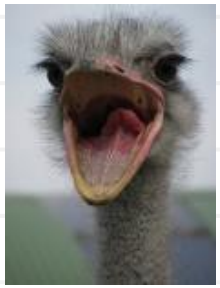
$$\max_r -\lambda \|r\| + \Phi_{\text{Last}}(x+r)[y']$$

[Szegedy et al 2014]

Generating adversarial perturbations







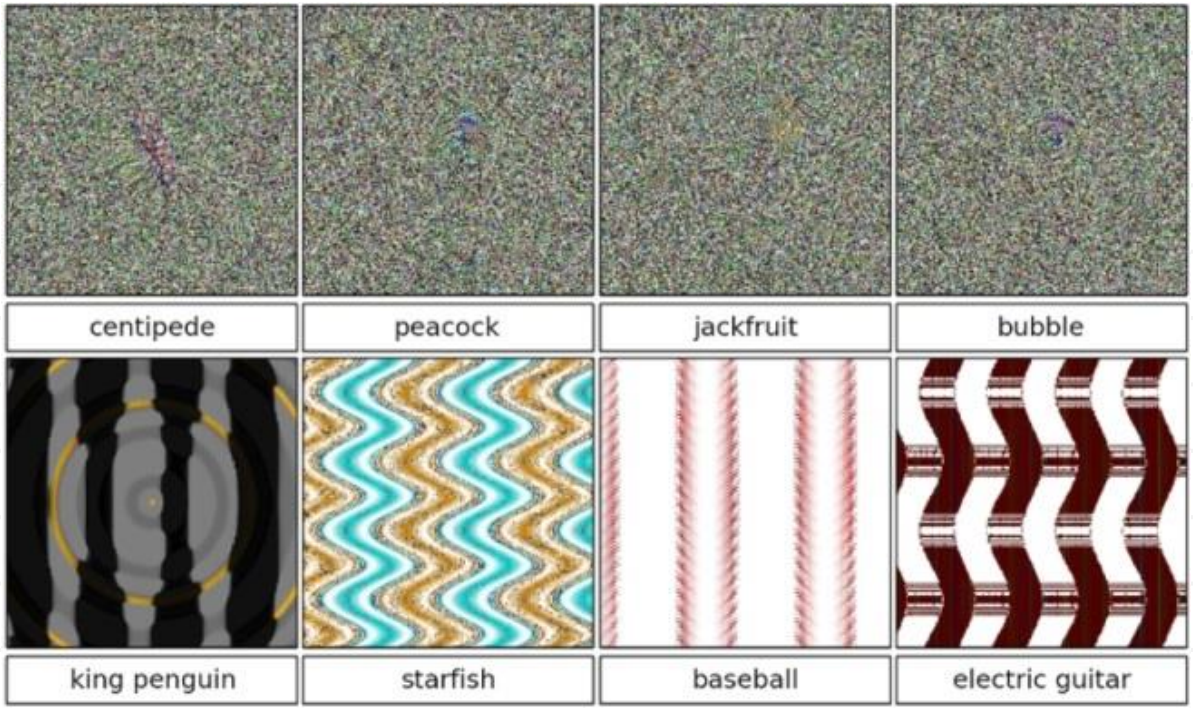
[Szegedy et al 2014]

Biological neural networks can be fooled too!

Some illusions are universal

[Elsayed et al. ArXiV18]

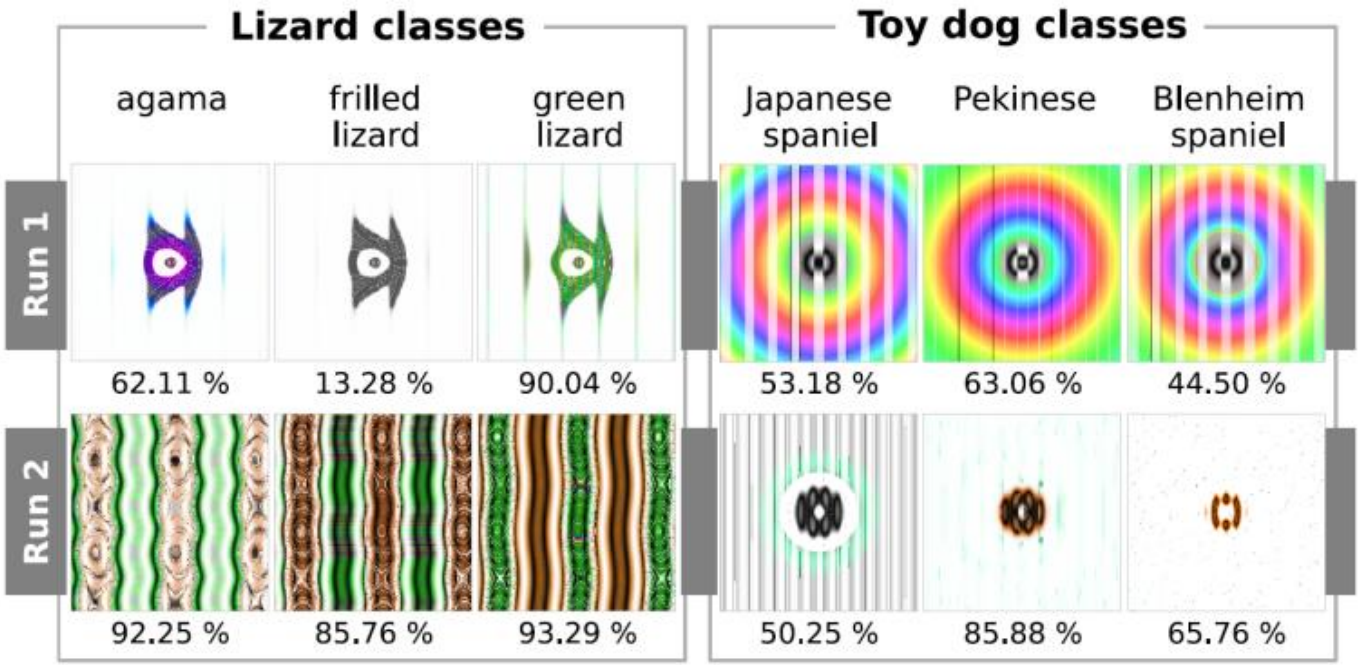
High-confidense patterns

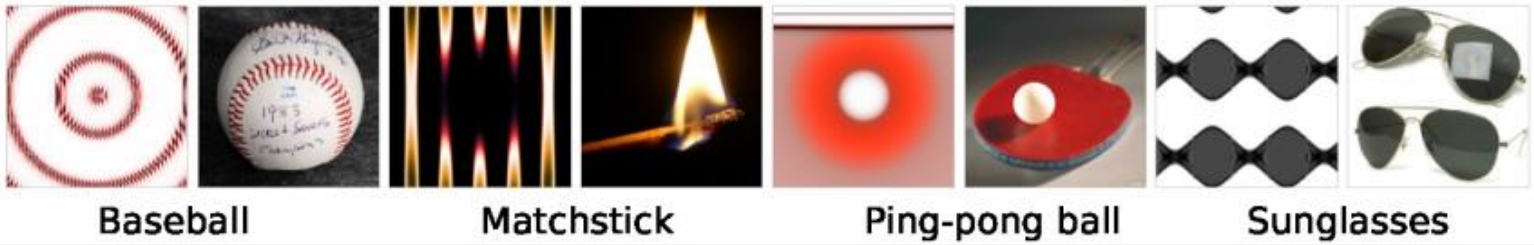


[Nguyen et al. CVPR15]:

- Evolutionary "breeding" of high-confidence patterns
- Compositional pattern-producing networks (CPPN) – superposition of sines, Gaussians, polynoms

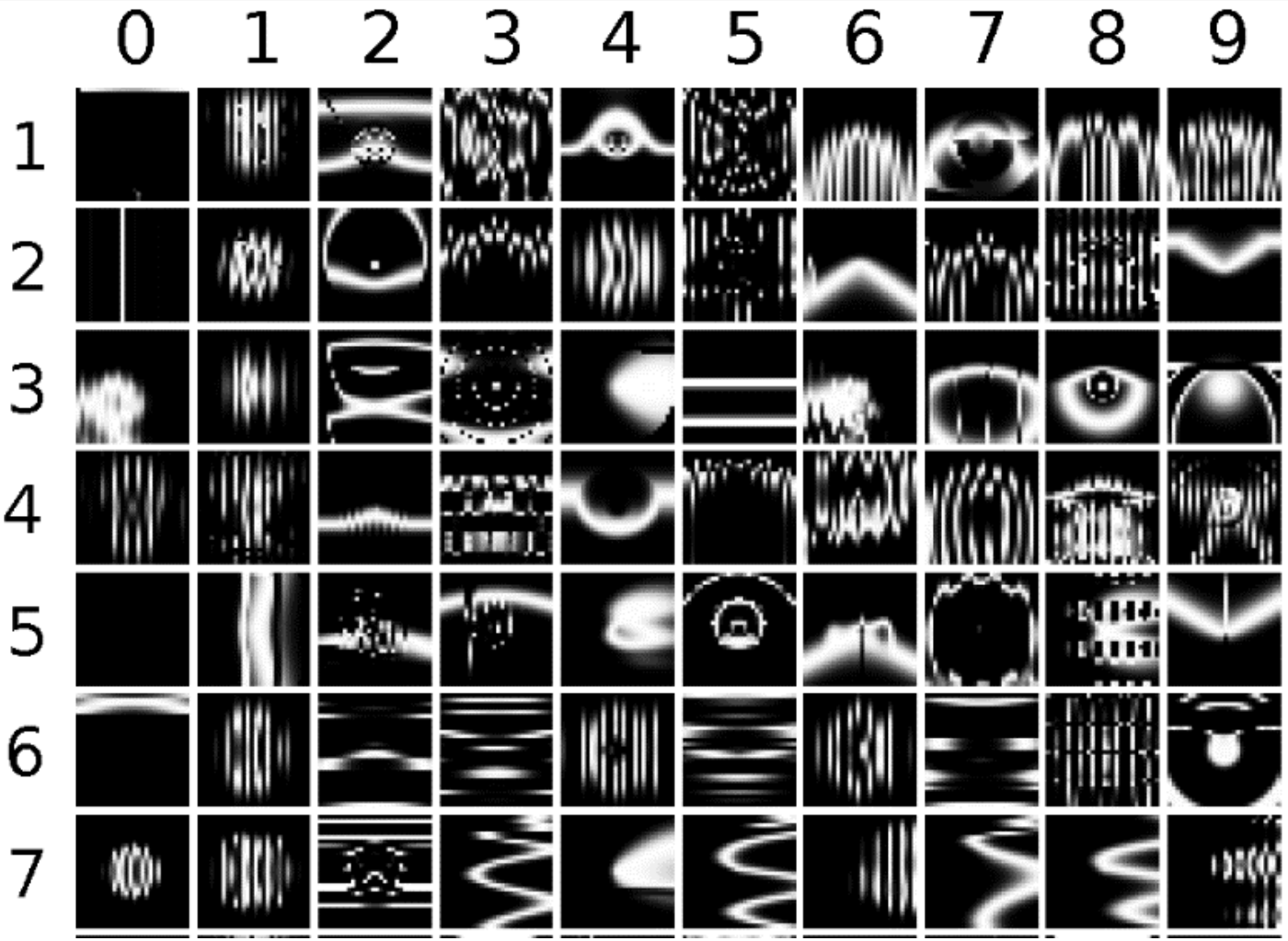
High-confidense patterns





[Nguyen et al. CVPR15]

Adversarial examples do not help



[Nguyen et al. CVPR15]

ConvNets can learn anything

[Zhang et al. ICLR17]: **typical ConvNet architectures can fit random labels on ImageNet**

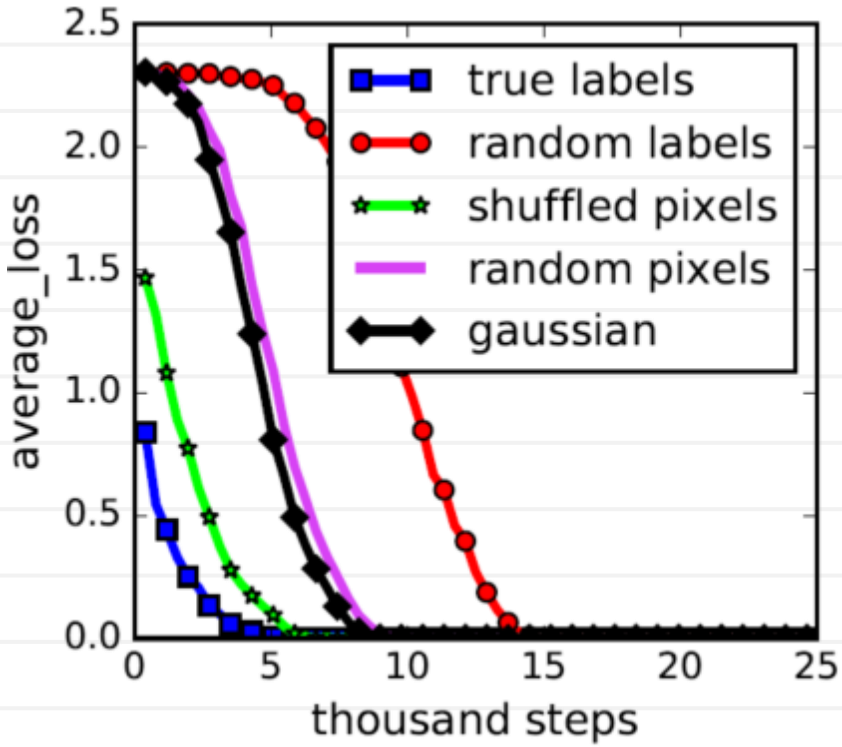
(and regularization tricks do not help much)

data aug	dropout	weight decay	top-1 train	top-5 train	top-1 test	top-5 test
ImageNet 1000 classes with the original labels						
yes	yes	yes	92.18	99.21	77.84	93.92
yes	no	no	92.33	99.17	72.95	90.43
no	no	yes	90.60	100.0	67.18 (72.57)	86.44 (91.31)
no	no	no	99.53	100.0	59.80 (63.16)	80.38 (84.49)
Alexnet (Krizhevsky et al., 2012)			-	-	-	83.6
ImageNet 1000 classes with random labels						
no	yes	yes	91.18	97.95	0.09	0.49
no	no	yes	87.81	96.15	0.12	0.50
no	no	no	95.20	99.14	0.11	0.56

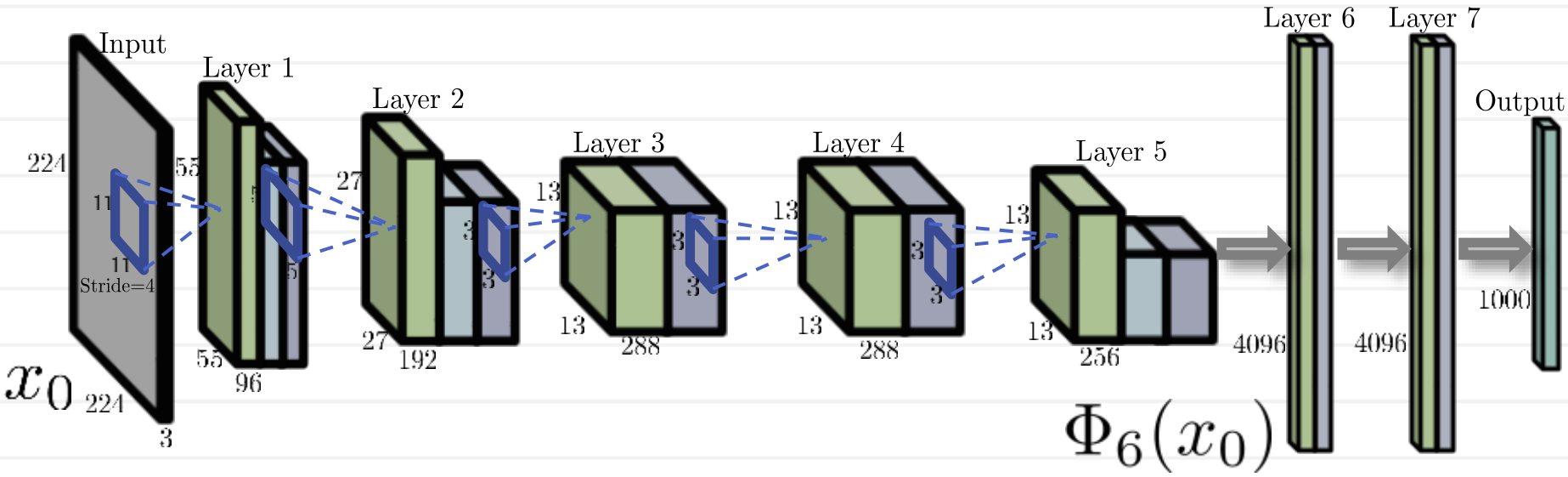
ConvNets can learn anything

[Zhang et al. ICLR17]: **typical ConvNet architectures can fit random labels on CIFAR perfectly!**

(and regularization tricks do not help much!)



Generating images by Inverting CNNs



$$\hat{x} = \arg\min_x \|\Phi_6(x) - \Phi_6(x_0)\|^2 + \lambda R(x)$$

Standard regularizer:

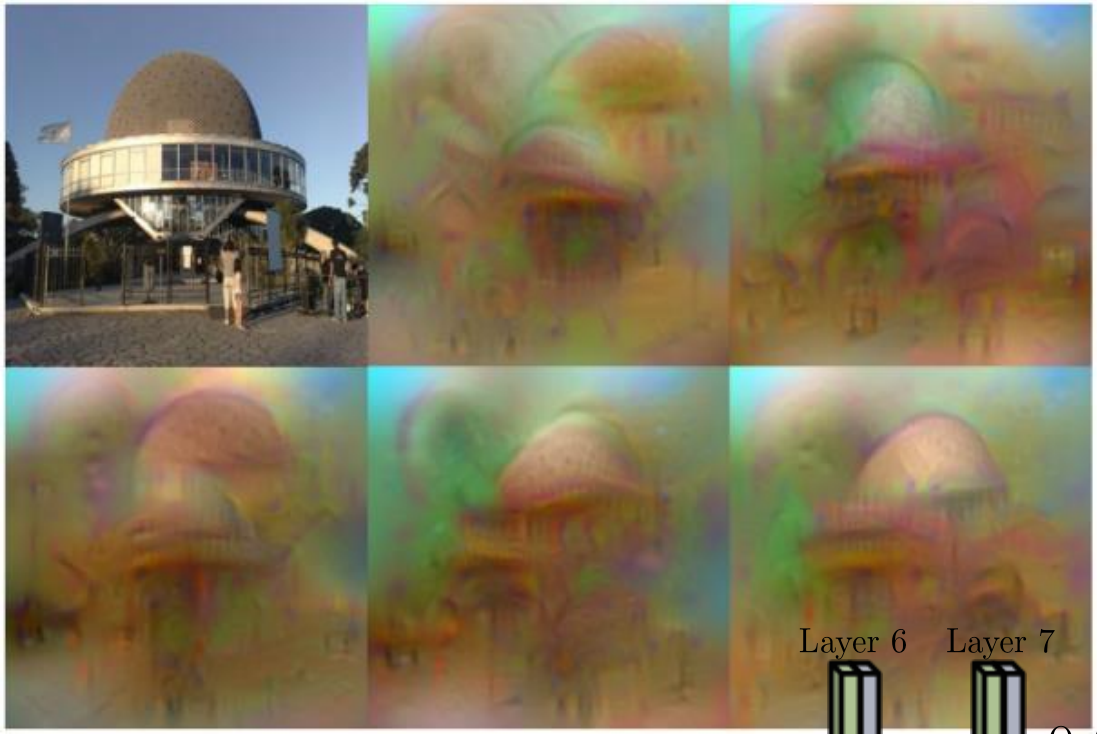
$$R(x) = \sum_{p,q} \left((x_{p,q+1} - x_{p,q})^2 + (x_{p+1,q} - x_{p,q})^2 \right)^{\beta/2}$$

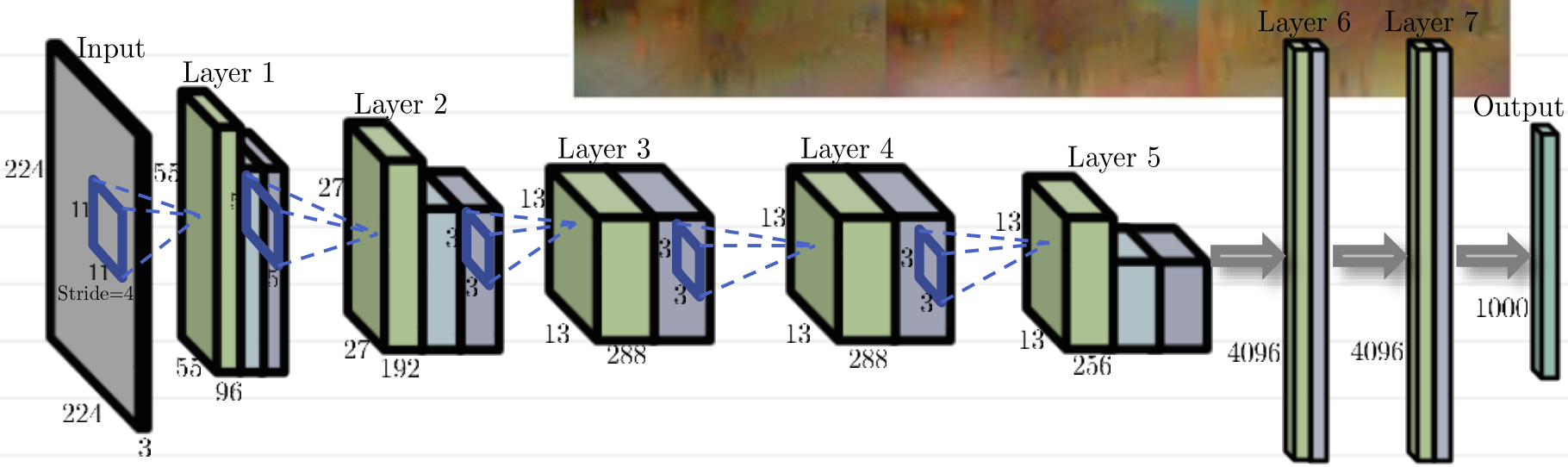
[Mahendrahan & Vedaldi CVPR15]

Generating images by Inverting CNNs

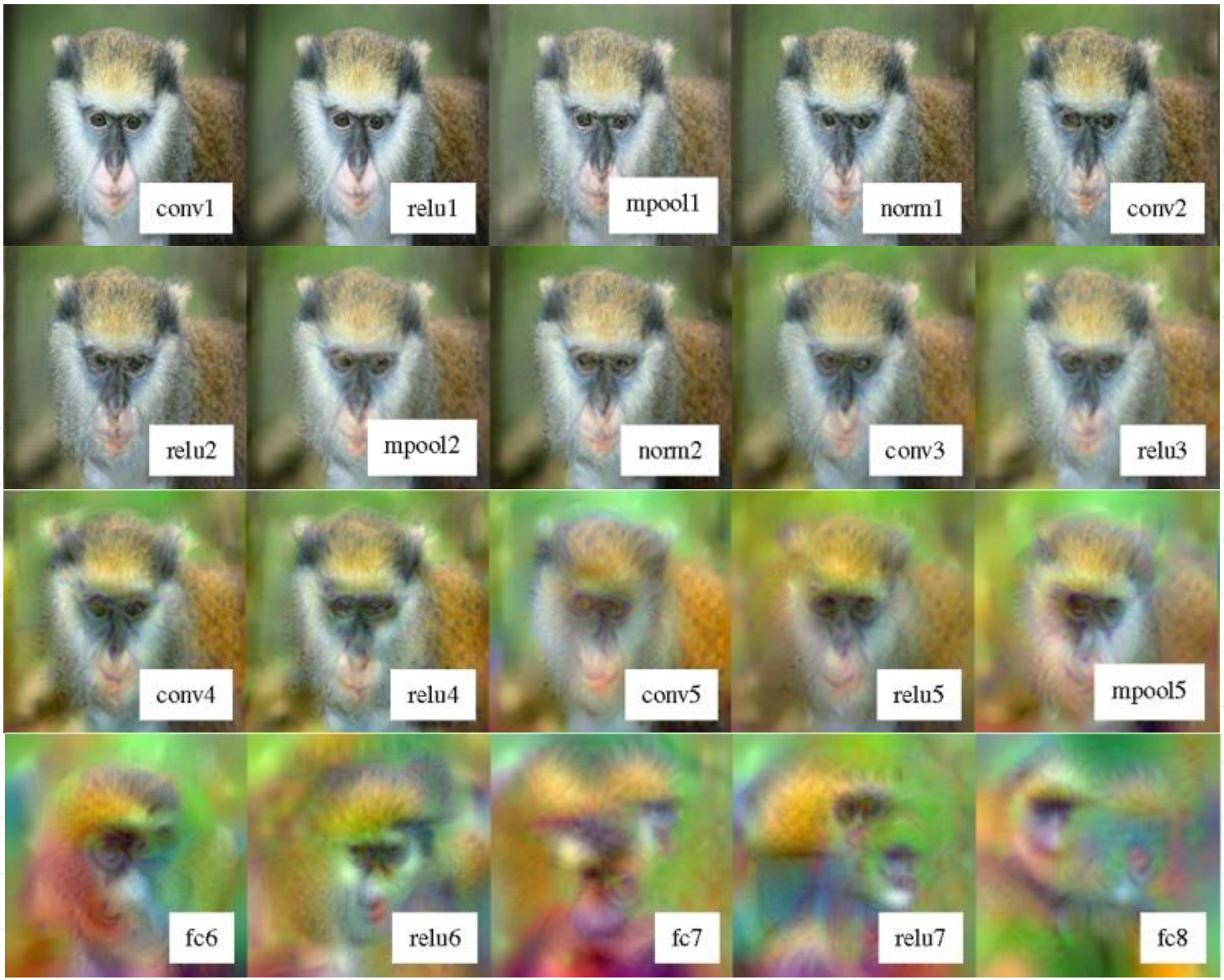
$$\hat{x} = \arg\min_x \|\Phi_6(x) - \Phi_6(x_0)\|^2 + \lambda R(x)$$

[Mahendrahan & Vedaldi CVPR15]





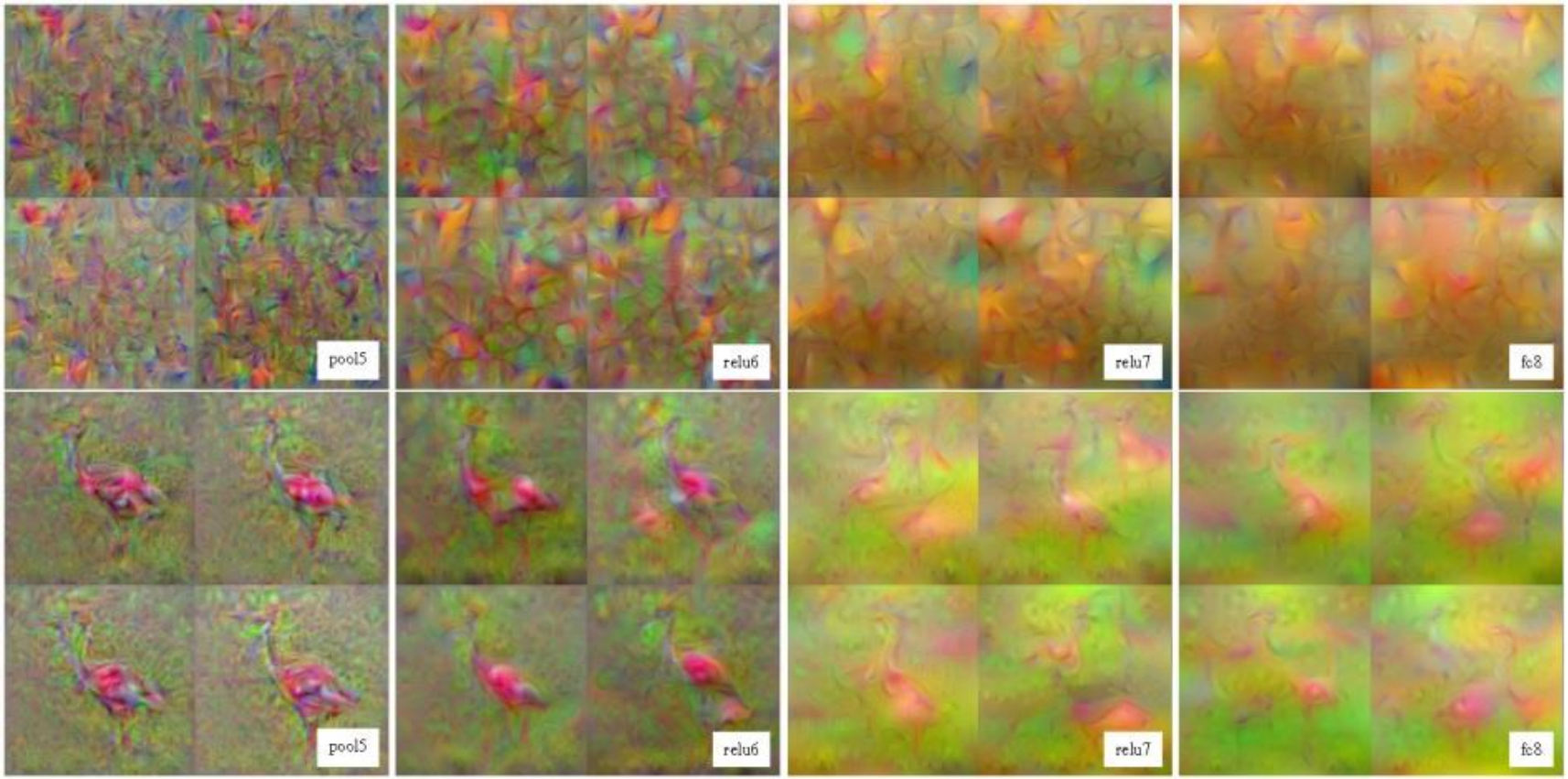
Generating images by Inverting CNNs



[Mahendrahan & Vedaldi CVPR15]

$$\hat{x} = \arg\min_{x} \|\Phi_6(x) - \Phi_6(x_0)\|^2 + \lambda R(x)$$

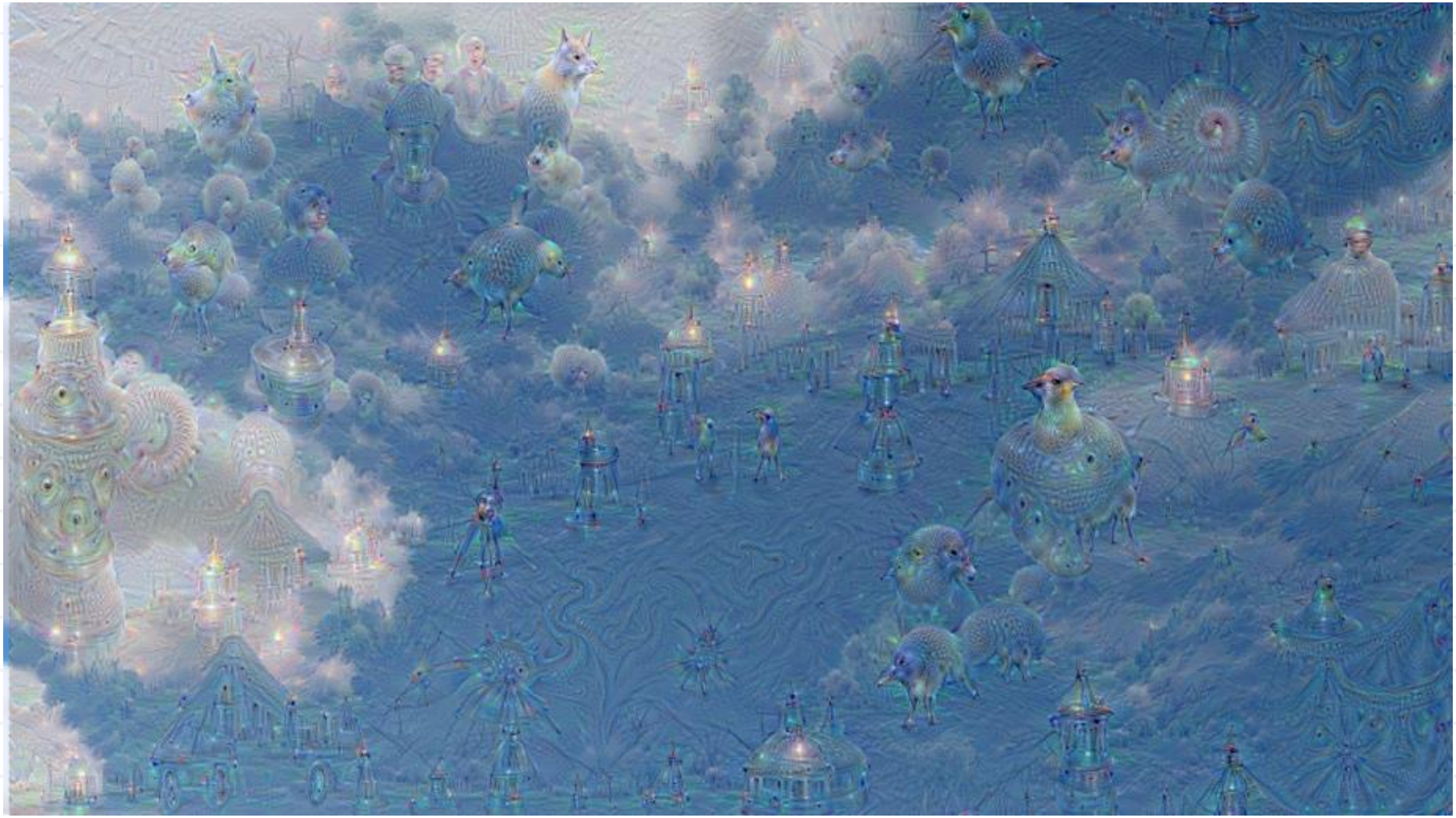
Multiple restarts



[Mahendrahan & Vedaldi CVPR15]

Inceptionism: DeepDream

ConvNet on drugs: $\max_x \|\Phi_k(x)\| - R(x)$

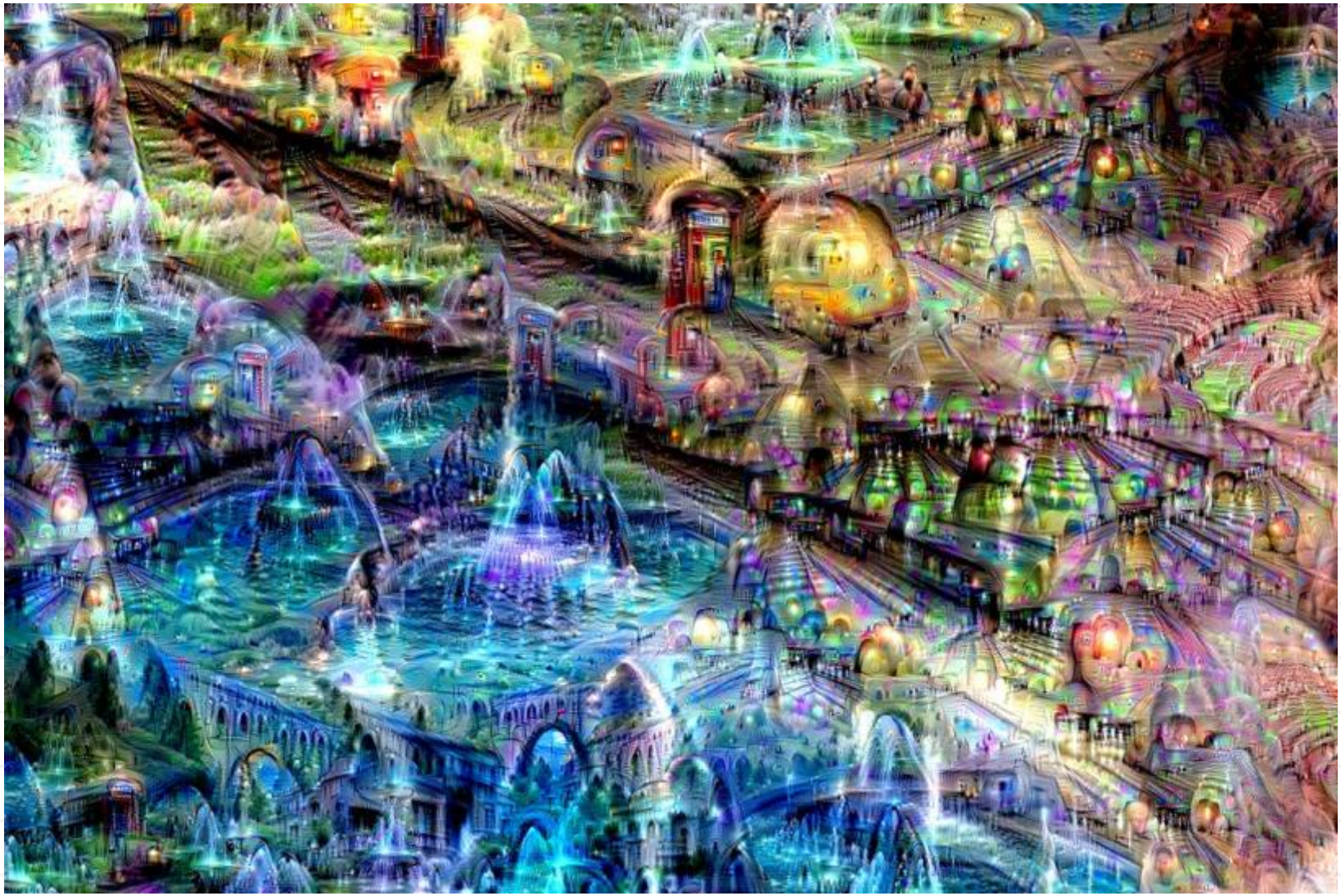


Inceptionism: DeepDream

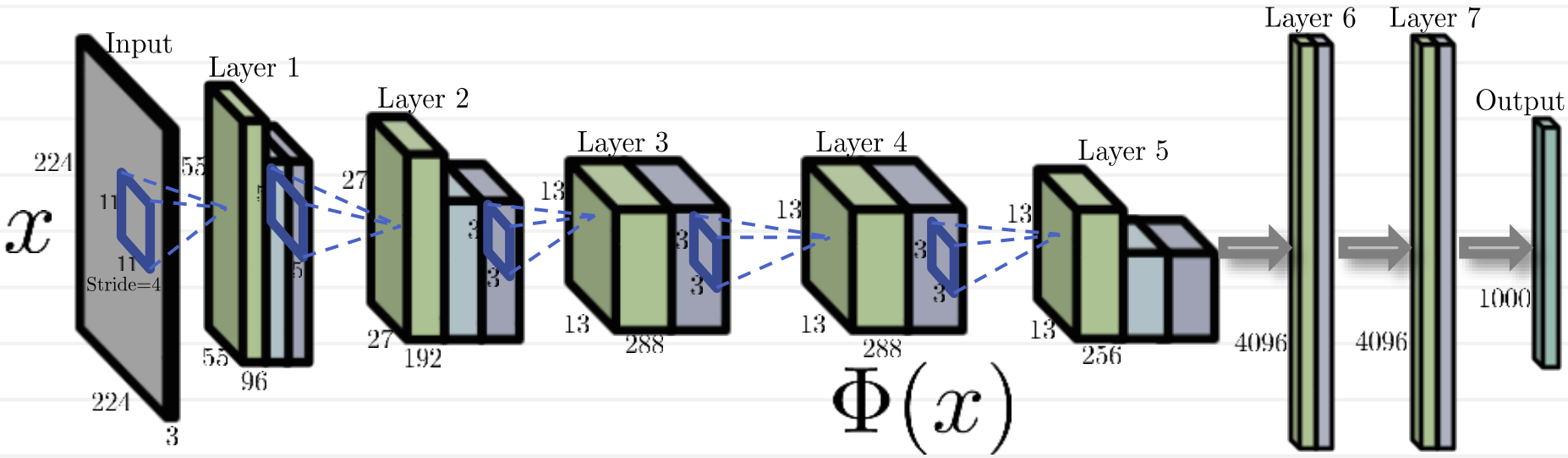
ConvNet on drugs: $\max_x \|\Phi_k(x)\| - R(x)$

Inceptionism: DeepDream

ConvNet on drugs (adding zoom-ins):



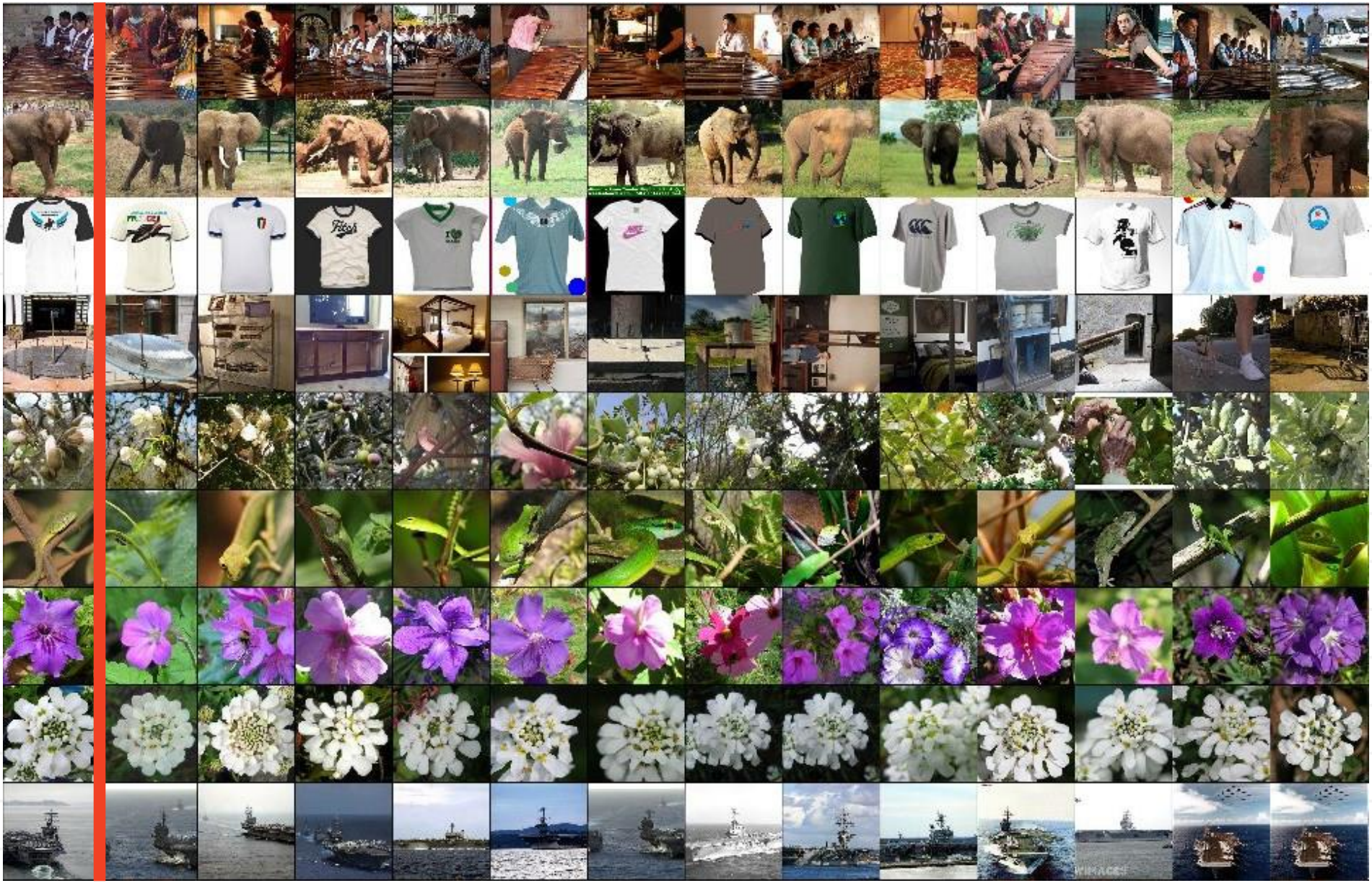
Representations inside the neural network



Lots of important questions:

- What are their properties?
- Are they redundant?
- Are they invertible?
- Are the intermediate representations useful?

Retrieval using learned representations



[Krizhevsky et al. NIPS12]

Retrieving same objects/buildings

Query:

[Babenko et al. 2014]
comparing representations:

Retrieving same objects/buildings

Query:

[Babenko et al. 2014]
comparing representations:

Compact descriptors

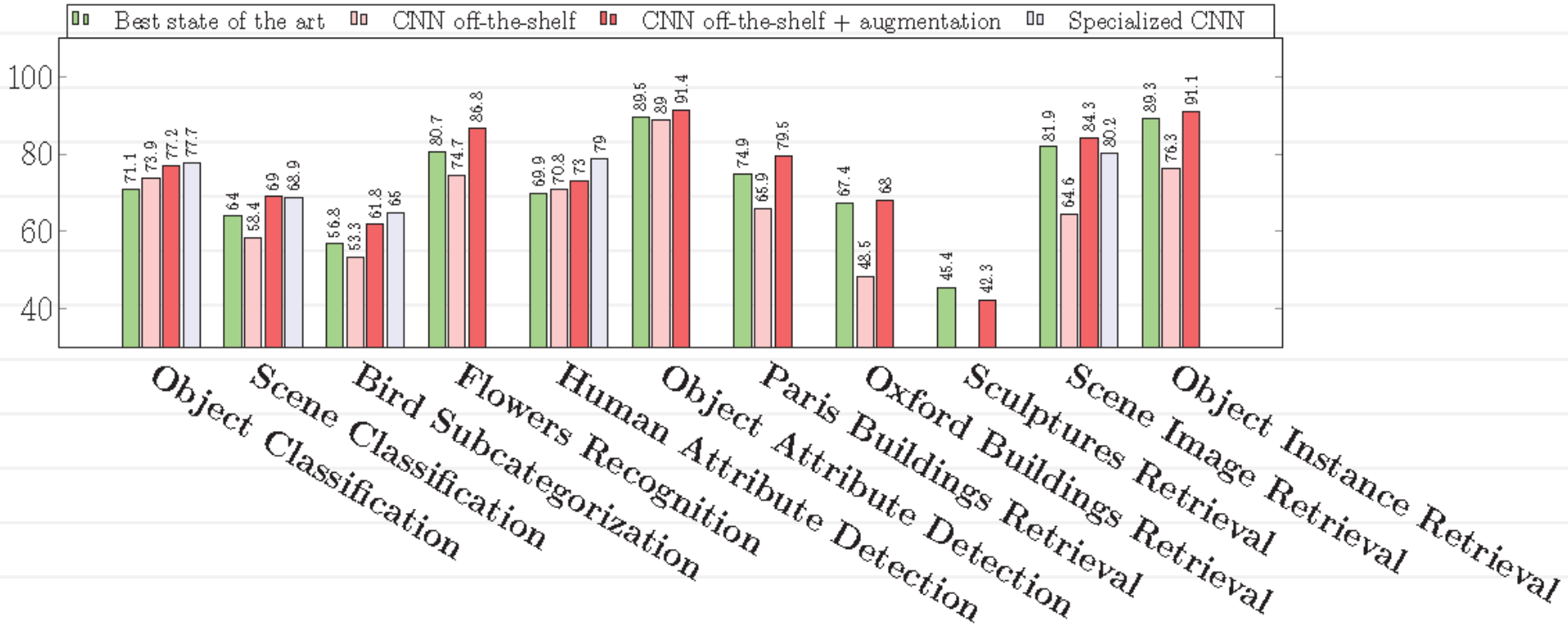
- Original descriptors can be compressed very well (e.g. to 128 dims)
- Even better representation for retrieval ([Babenko et al. ICCV2015]):

$$\Psi(x) = \sum_{p,q} \Phi_{LC}(x)[p,q]$$

$$\Psi'(x) = \mathrm{diag}(s_1, \ldots, s_N)^{-1} M_{\mathrm{PCA}} \Psi(x)$$

$$\Psi''(x) = \frac{\Psi'(x)}{\|\Psi'(x)\|}$$

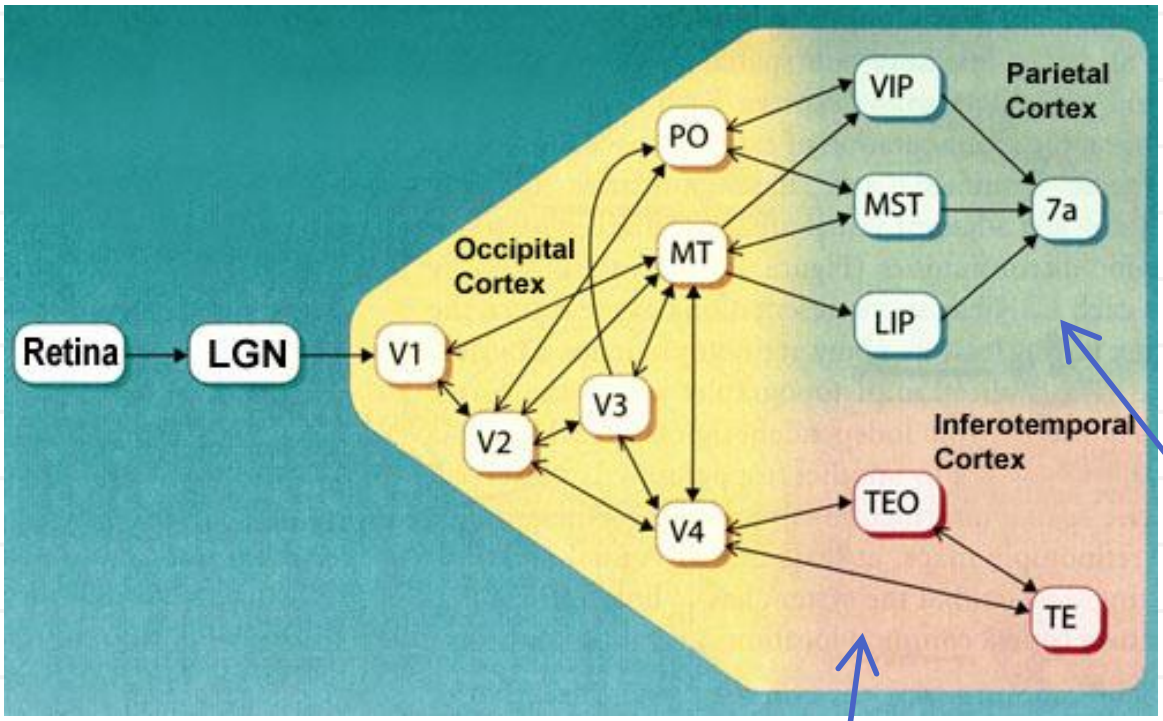
Deep features as generic representation

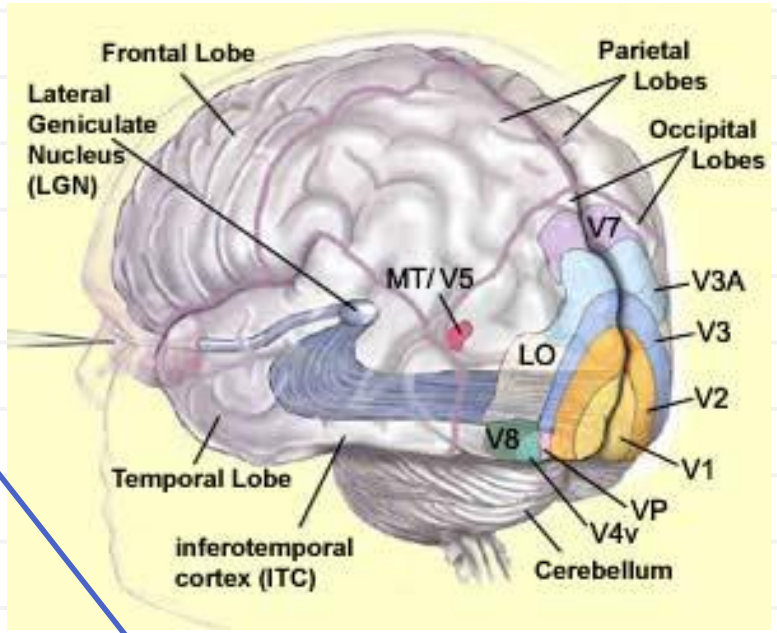


[Razavian et al. 2014]: ImageNet-trained network -> representations-> linear SVM or L2 NN-search

Postprocessing: L2-normalize+PCA+ whitening+L2-normalize+power (all standard things used to postprocess shallow features)

Visual cortex





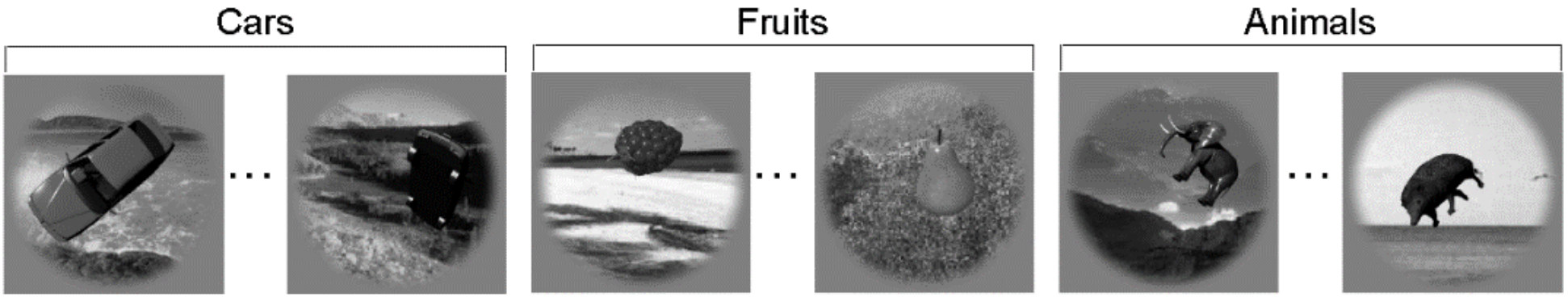
What? (objects, faces)

Where? (localizations,motions, actions)

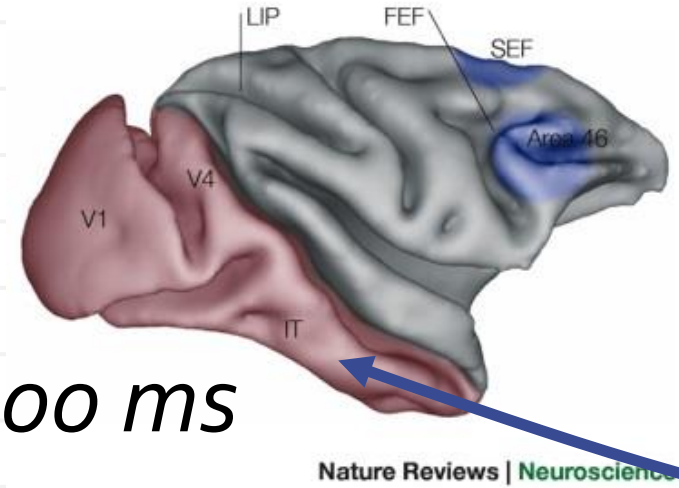
Source: "The brain from top to bottom" website

Human vs machine

[Cadieu et al. 2014]



7 generic classes



100 ms

vs.

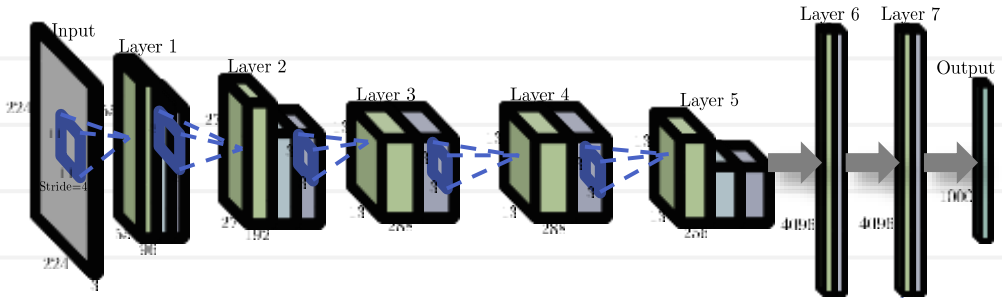


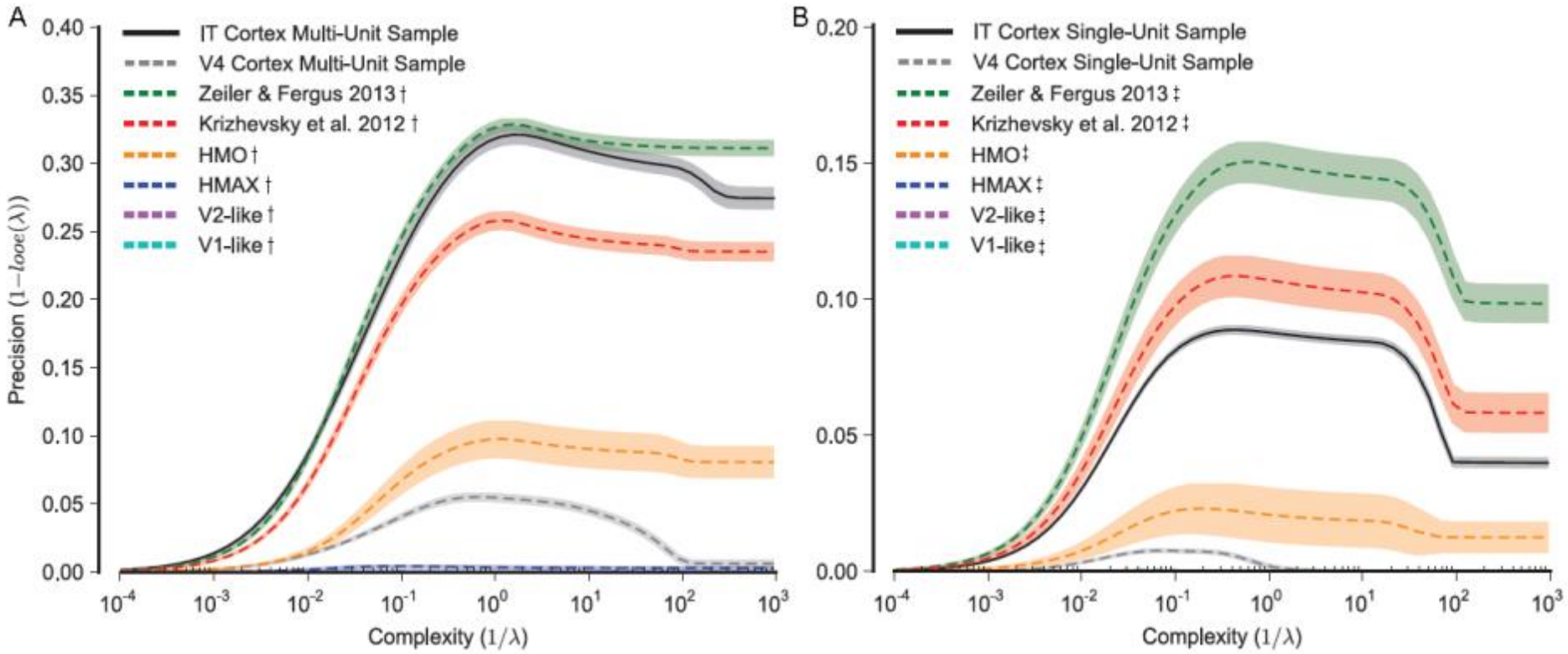
Image: [Sugrue et al. 2005]

Human vs machine

[Cadieu et al. 2014]

80 ("multi-unit") vs 80

40 (single neurons) vs 40

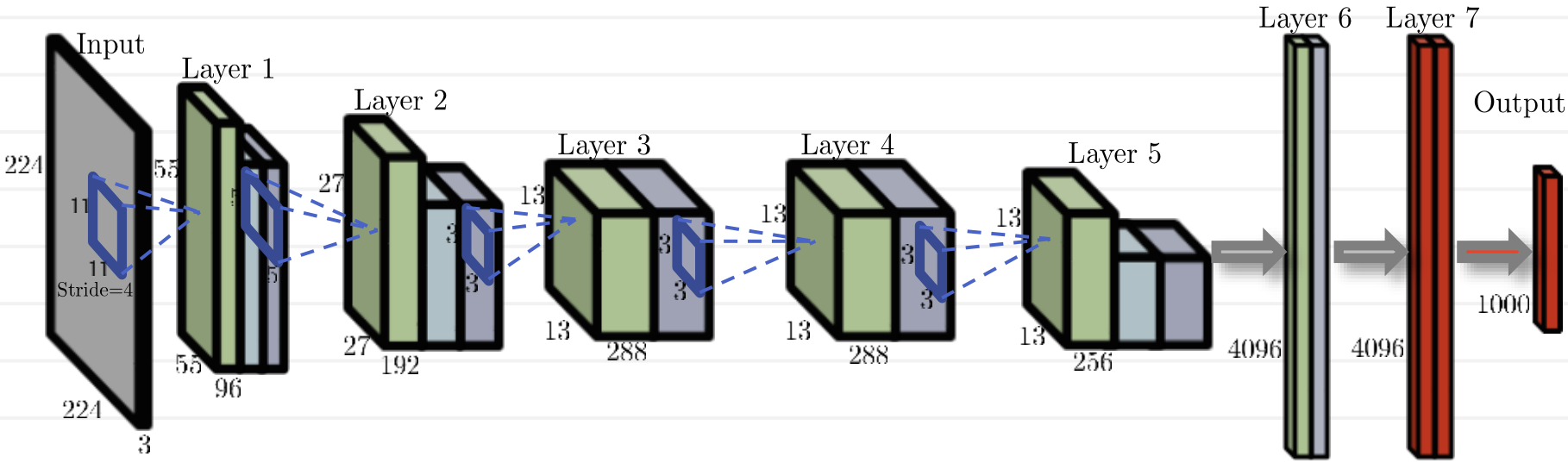


- Macaques were fixating on the images
- 100ms display, spiking rate recorded after that
- Averaging over 6 displays for each image/10 image crops

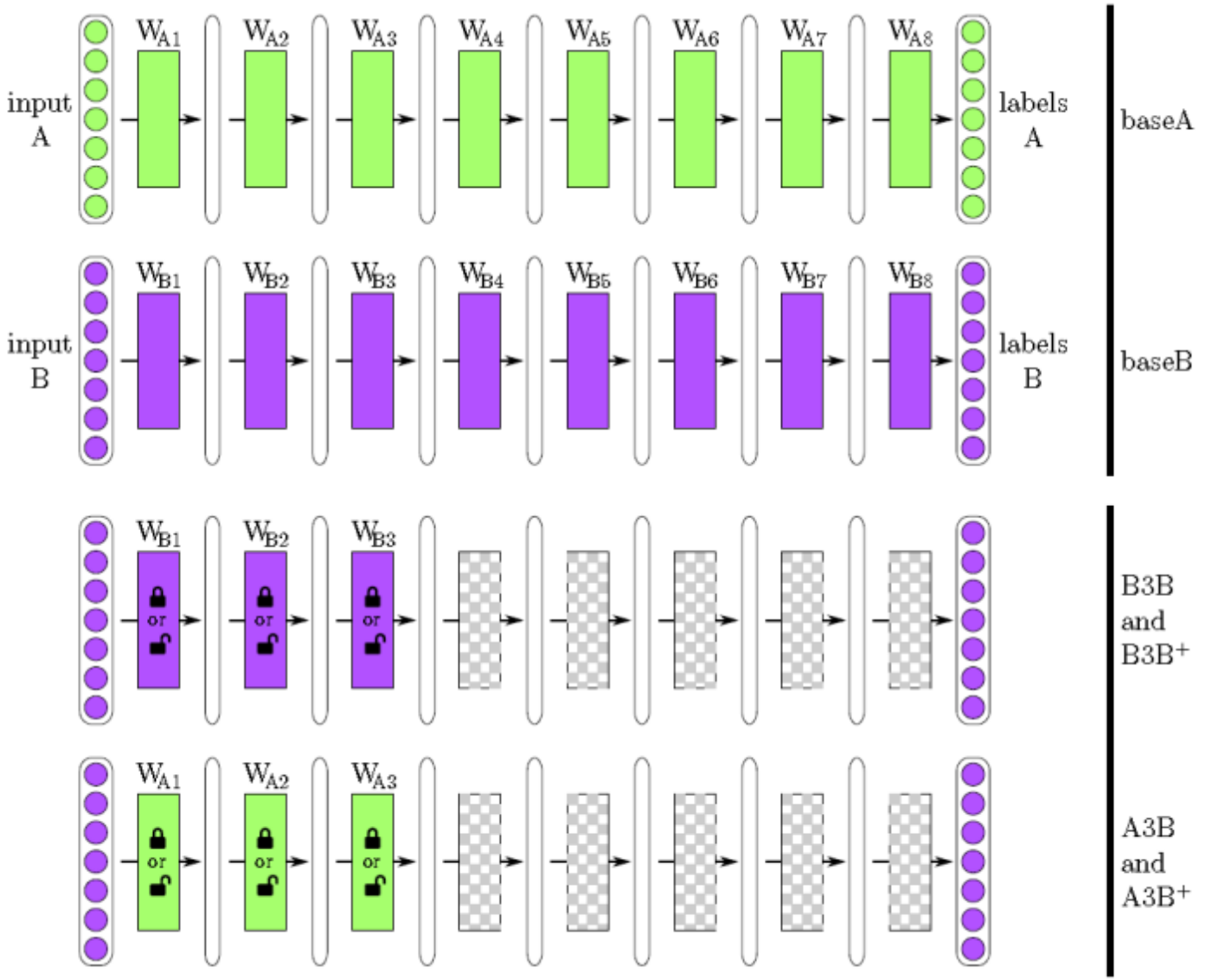
Fine-tuning

Lots of papers (e.g. [Oquab et al. CVPR 14], [Babenko et al. ECCV 14], [Yosinski et al. NIPS14]) "fine-tune" the network on smaller datasets:

- Chop off top layers
- Initialize new top layers
- Run SGD

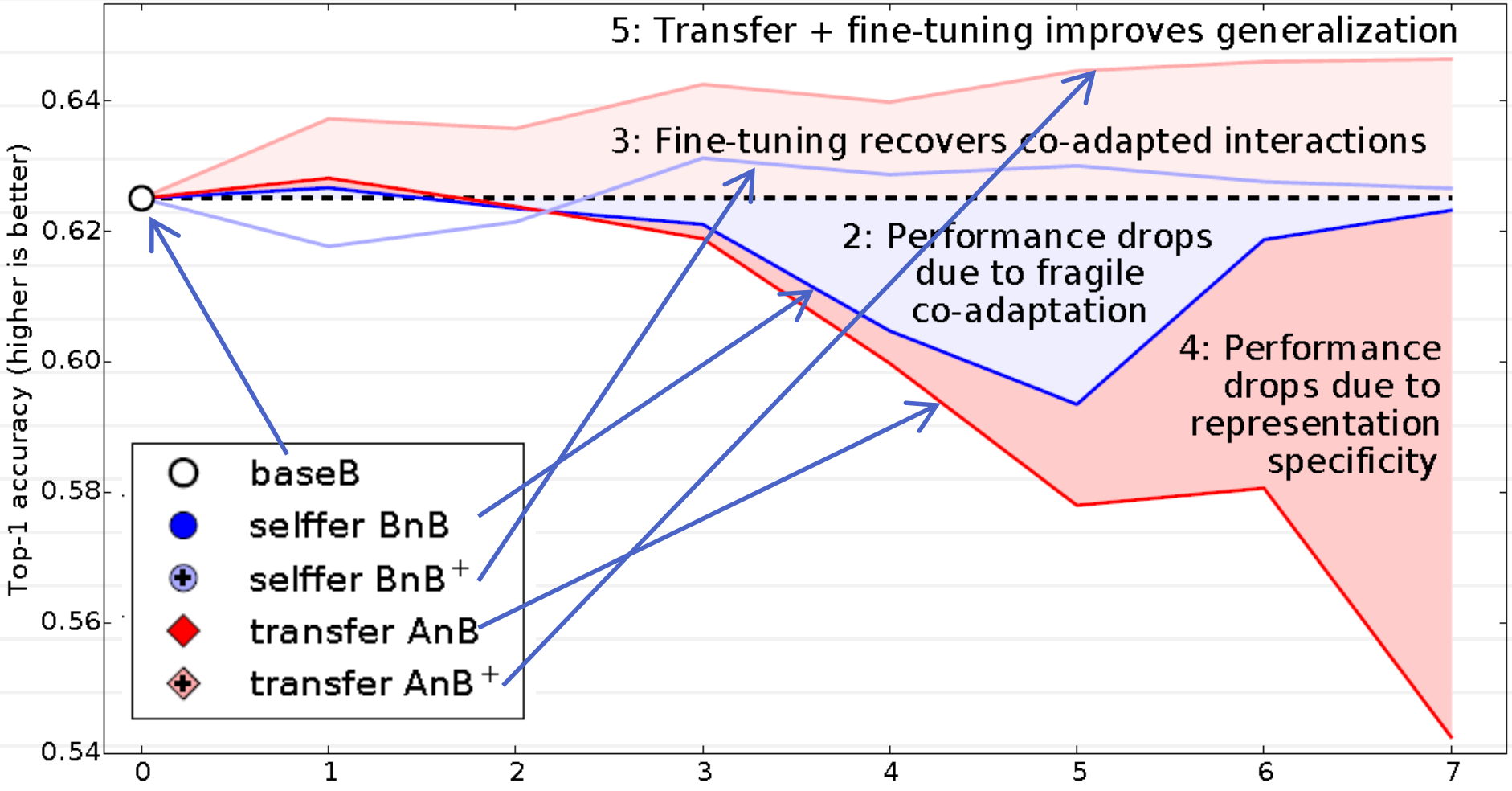


How to fine-tune



[Yosinski et al. 2014]: two halves of Image-NET

How to fine-tune?



"+" = do not lock bottom layers

[Yosinski et al. NIPS2014]

Fine-tuning helps retrieval

- Image-Net -> "Landmarks"

Oxford Buildings:
0.388->0.523
INRIA Holidays:
0.727->0.769



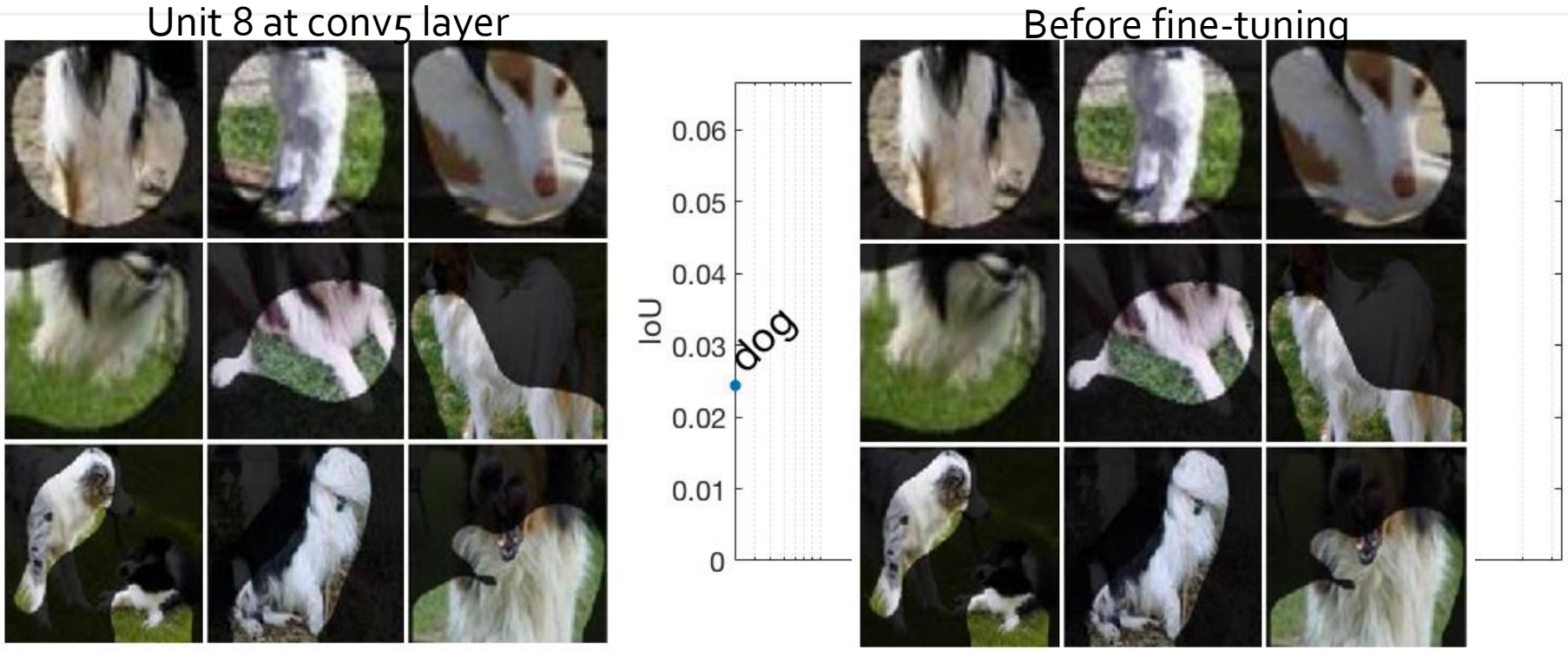
[Babenko et al. 2014]

Fine-tuning examples

[Babenko et al. 2014]

Fine-tuning from ImageNet to Places

[Zhou et al. CVPR17]



Fine-tuning from Places to ImageNet

[Zhou et al. CVPR17]

Unit 35 at conv5 layer

Before fine-tuning

Applying CNN in practice

New problem

- PyTorch torchvision
- Caffe zoo
- MatConvNet zoo
- Lasagne Recipes

Literature

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun:
Deep Residual Learning for Image Recognition. CoRR abs/1512.03385 (2015)

Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton:
ImageNet Classification with Deep Convolutional Neural Networks. NIPS 2012

Karen Simonyan, Andrea Vedaldi, Andrew Zisserman:
Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps.CoRR abs/1312.6034 (2013)

Karen Simonyan, Andrew Zisserman:
Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR abs/1409.1556 (2014)

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan,Vincent Vanhoucke, Andrew Rabinovich:
Going deeper with convolutions. CVPR 2015: 1-9

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, Rob Fergus:
Intriguing properties of neural networks. CoRR abs/1312.6199 (2013)

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke:
Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. CoRRabs/1602.07261 (2016)

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, Fei-Fei Li:
ImageNet Large Scale Visual Recognition Challenge. CoRR abs/1409.0575 (2014)

Literature

Matthew D. Zeiler, Rob Fergus:
Visualizing and Understanding Convolutional Networks. ECCV (1) 2014: 818-833

Anh Mai Nguyen, Jason Yosinski, Jeff Clune:
Deep neural networks are easily fooled: High confidence predictions for unrecognizable images.CVPR 2015: 427-436

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, Stefan Carlsson:
CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. CVPR Workshops 2014: 512-519

Jason Yosinski, Jeff Clune, Yoshua Bengio, Hod Lipson:
How transferable are features in deep neural networks? NIPS 2014, 3320-3328

Maxime Oquab, Léon Bottou, Ivan Laptev, Josef Sivic:
Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks.CVPR 2014: 1717-1724

Charles F. Cadieu, Ha Hong, Daniel Yamins, Nicolas Pinto, Diego Ardila, Ethan A. Solomon, Najib J. Majaj,James J. DiCarlo:
Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. PLoS Computational Biology 10(12) (2014)

Artem Babenko, Victor S. Lempitsky:
Aggregating Local Deep Features for Image Retrieval. ICCV 2015: 1269-1277

Artem Babenko, Anton Slesarev, Alexander Chigorin, Victor S. Lempitsky:
Neural Codes for Image Retrieval. ECCV (1) 2014: 584-599

Literature

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, Oriol Vinyals:
Understanding deep learning requires rethinking generalization. CoRR abs/1611.03530 (2016)
[ICLR 2017]

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, Martin A. Riedmiller:
Striving for Simplicity: The All Convolutional Net. CoRR abs/1412.6806 (2014)

Bolei Zhou, David Bau, Aditya Khosla, Aude Oliva, Antonio Torralba:
Network Dissection: Quantifying Interpretability of Deep Visual Representations. CVPR 2017:
3319-3327

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann,
Wieland Brendel:
ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and
robustness. ICLR 2019