Visualizing ConvNets

Presenters:

Part 1 - Abhijit Sarkar

Part 2 – Tamoghna Roy

Slide credit: Vondrick et al.

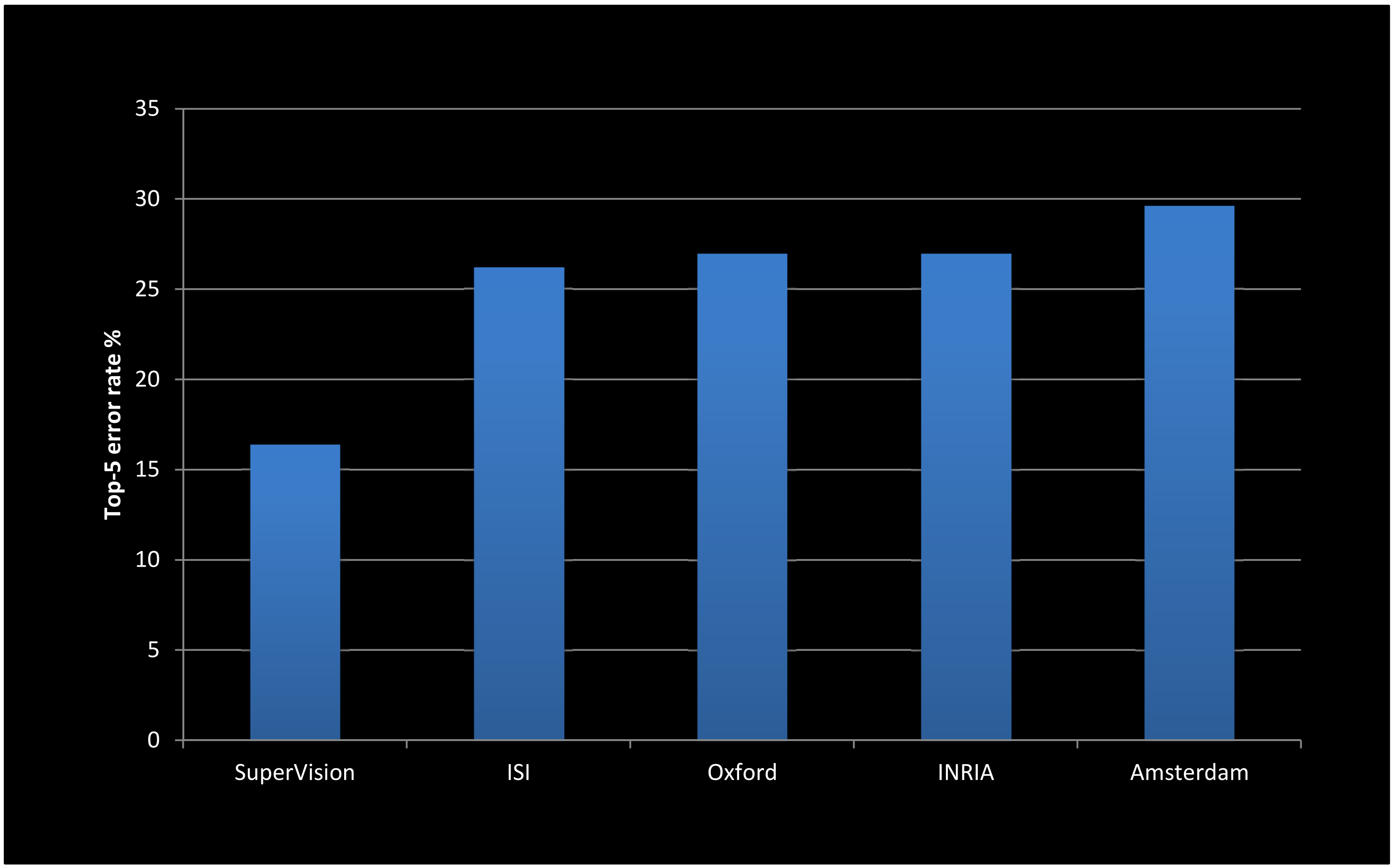
vs

Human Vision

HOG Vision

Slide credit: Vondrick et al.

ImageNet Challenge 2012

- Krizhevsky et al. -- **16.4% error** (top-5)
- 

Part 1

Visualizing and Understanding Convolutional Networks

Matthew D. Zeiler ZEILER@CS.NYU.EDU
Dept. of Computer Science, Courant Institute, New York University

Rob Fergus FERGUS@CS.NYU.EDU
Dept. of Computer Science, Courant Institute, New York University

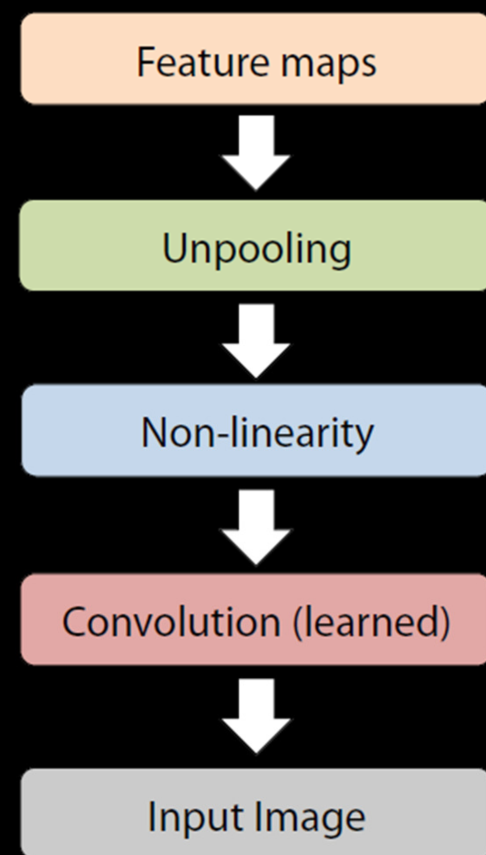
Topics

- Visualization of conv net
  - What are they learning over layers
  - How learning changes with epochs
  - Feature invariance
  - Occlusion experiment
  - Part based model
- New architecture and Imagenet competition
  - Change in model size, layers
- Model generalization

Deconvolutional Networks

[Zeiler et al. CVPR'10, ICCV'11]

- Provides way to map activations at high layers back to the input
- Same operations as Convnet, but in reverse:
  - Unpool feature maps
  - Convolve unpooled maps
    - Filters copied from Convnet
- Used here purely as a probe
  - Originally proposed as unsupervised learning method
  - No inference, no learning



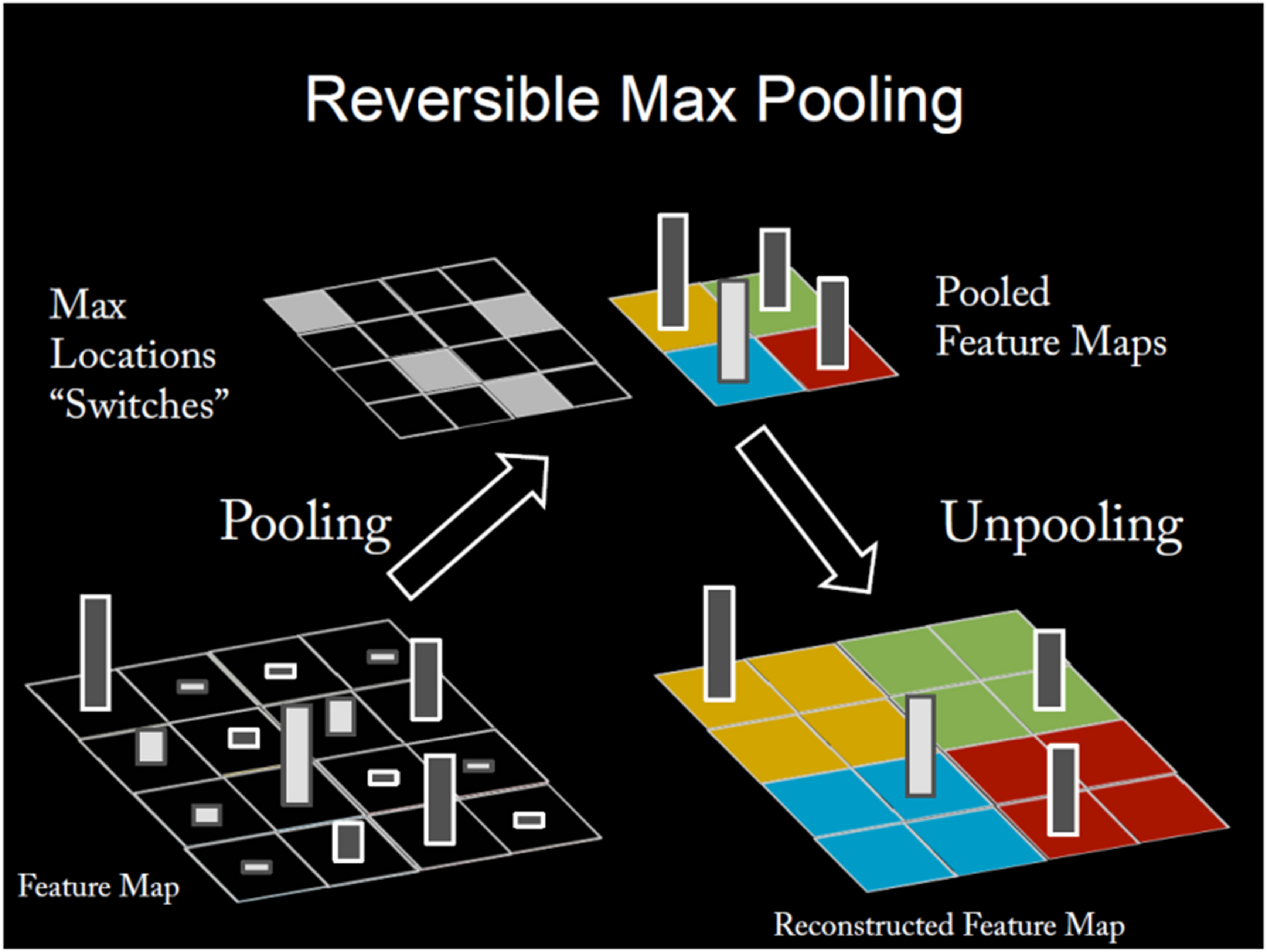
Slide credit: Mathew Zeiler

$$C = \frac{\lambda}{2} \| \sum_{k=1}^{K} z_k \oplus f_k - y \|_2^2 + \sum_{k=1}^{K} |z_k|_1$$

y = Input, z = Feature maps, f = Filters

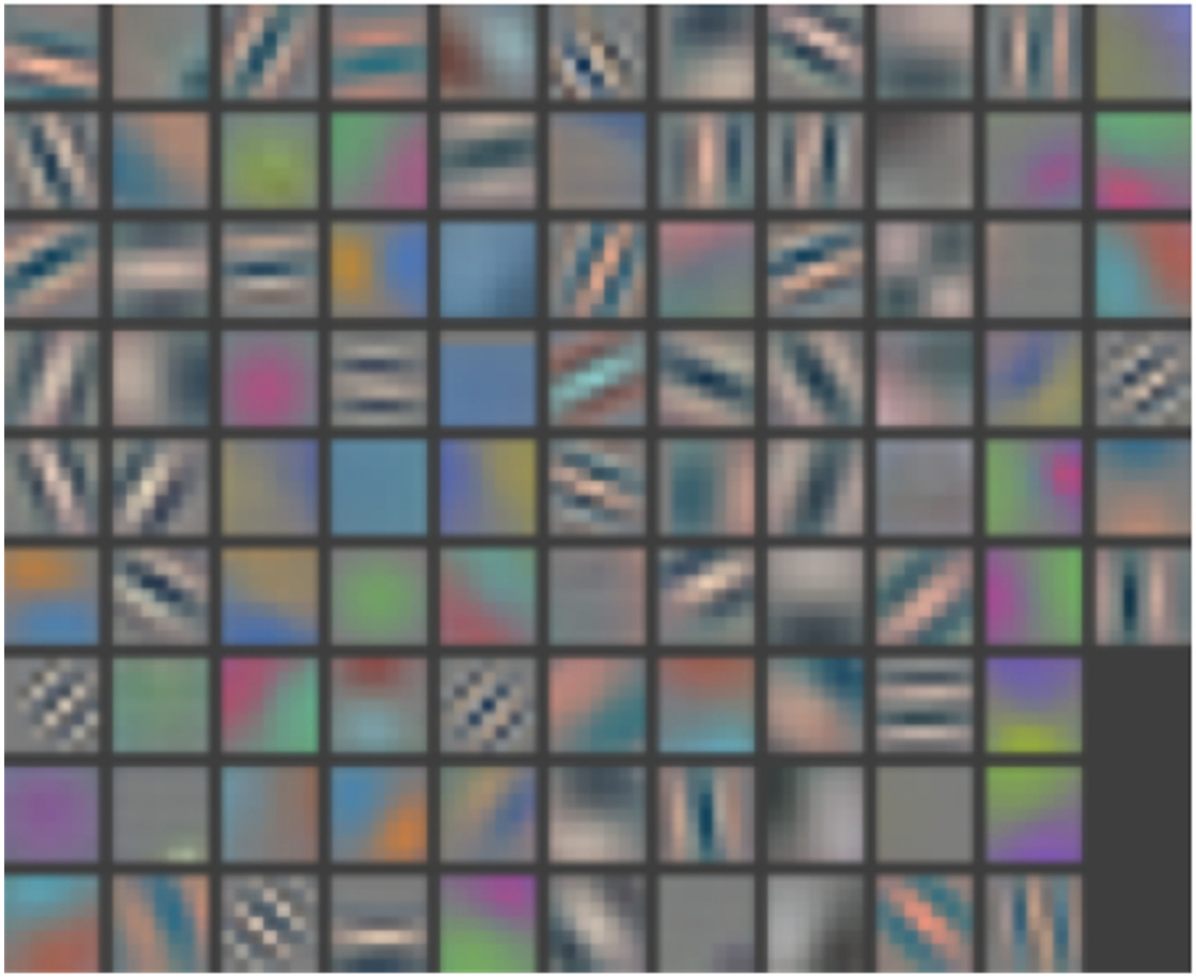
)

Slide credit: Mathew Zeiler

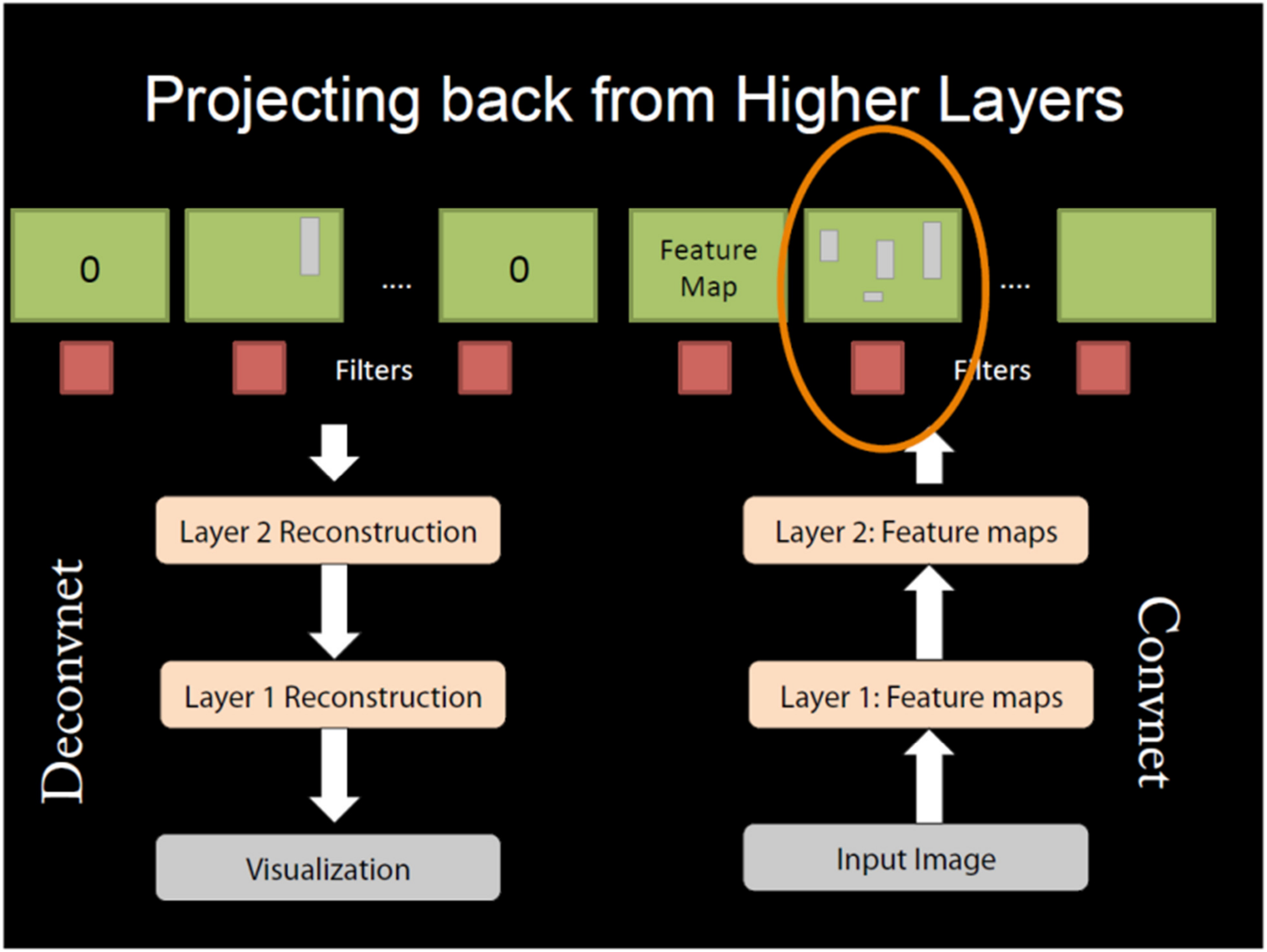


Slide credit: Mathew Zeiler

Layer 1 Filters



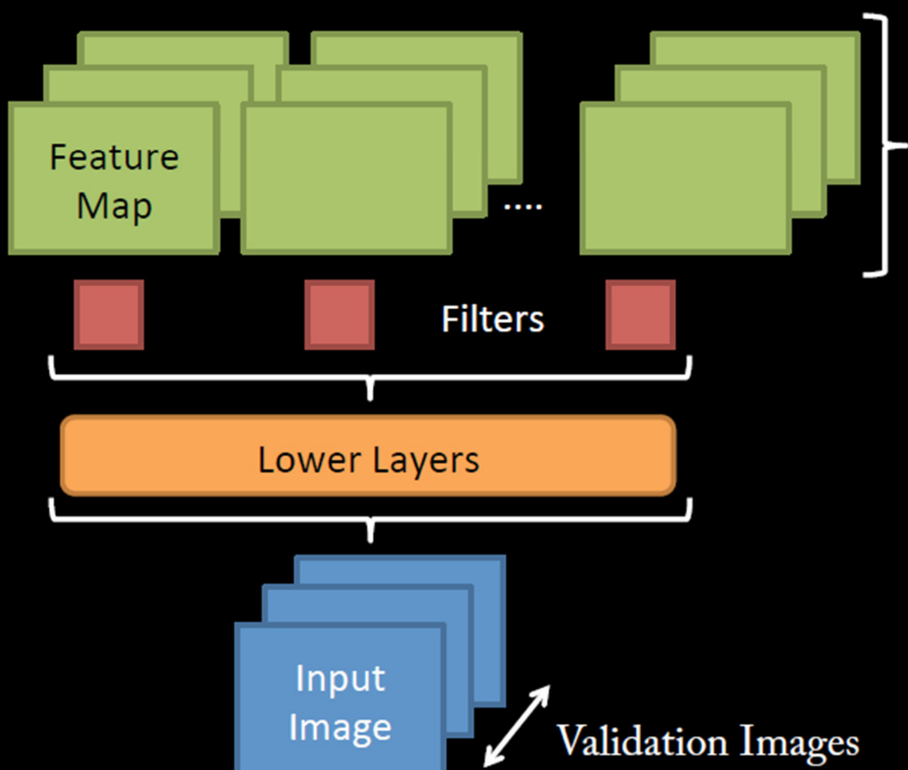
Slide credit: Mathew Zeiler



Slide credit: Mathew Zeiler

Visualizations of Higher Layers

- Use ImageNet 2012 validation set
- Push each image through network



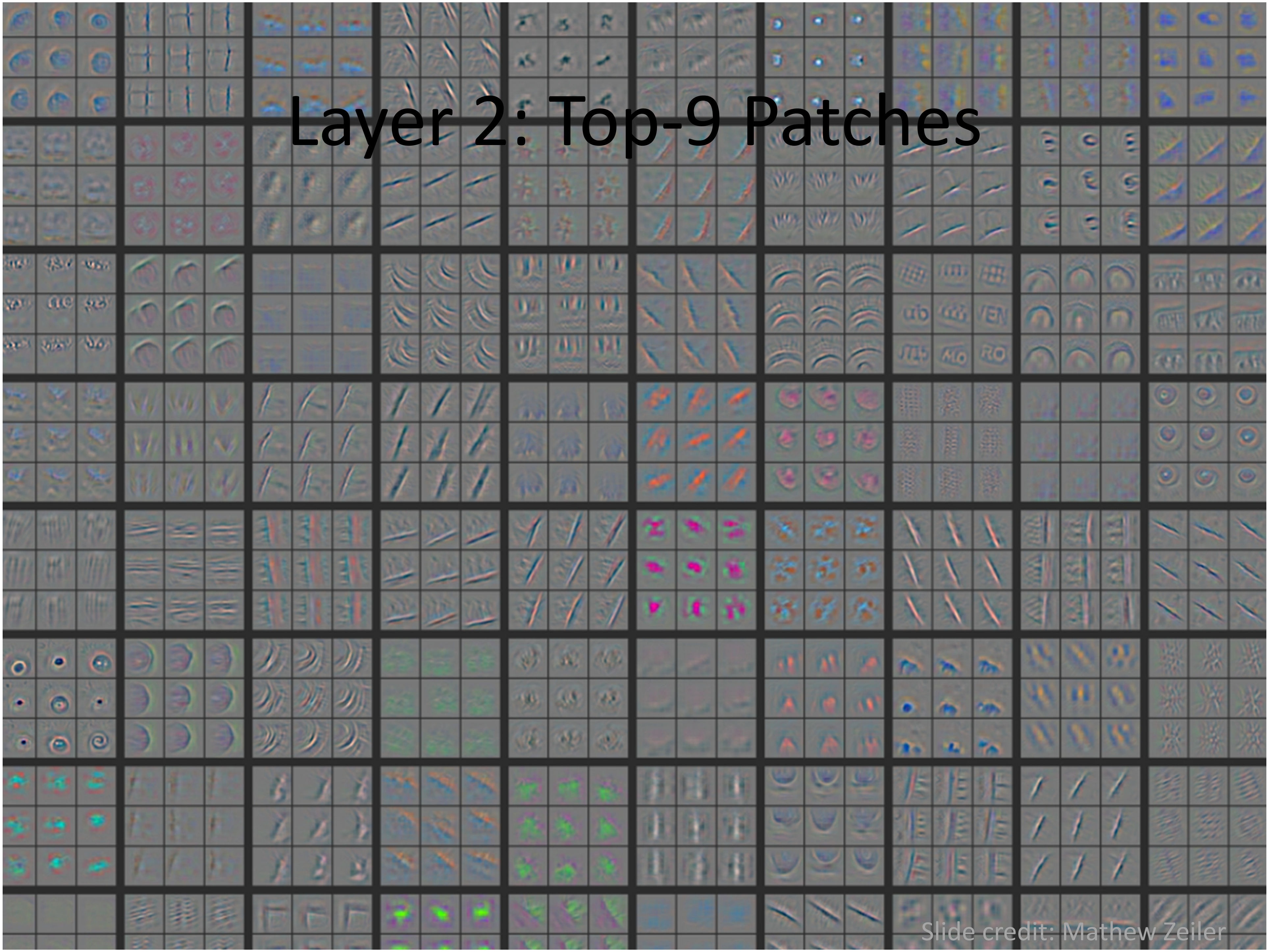
- Take max activation from feature map associated with each filter
- Use Deconvnet to project back to pixel space
- Use pooling “switches” peculiar to that activation

Slide credit: Mathew Zeiler

Layer 1: Top-9 Patches

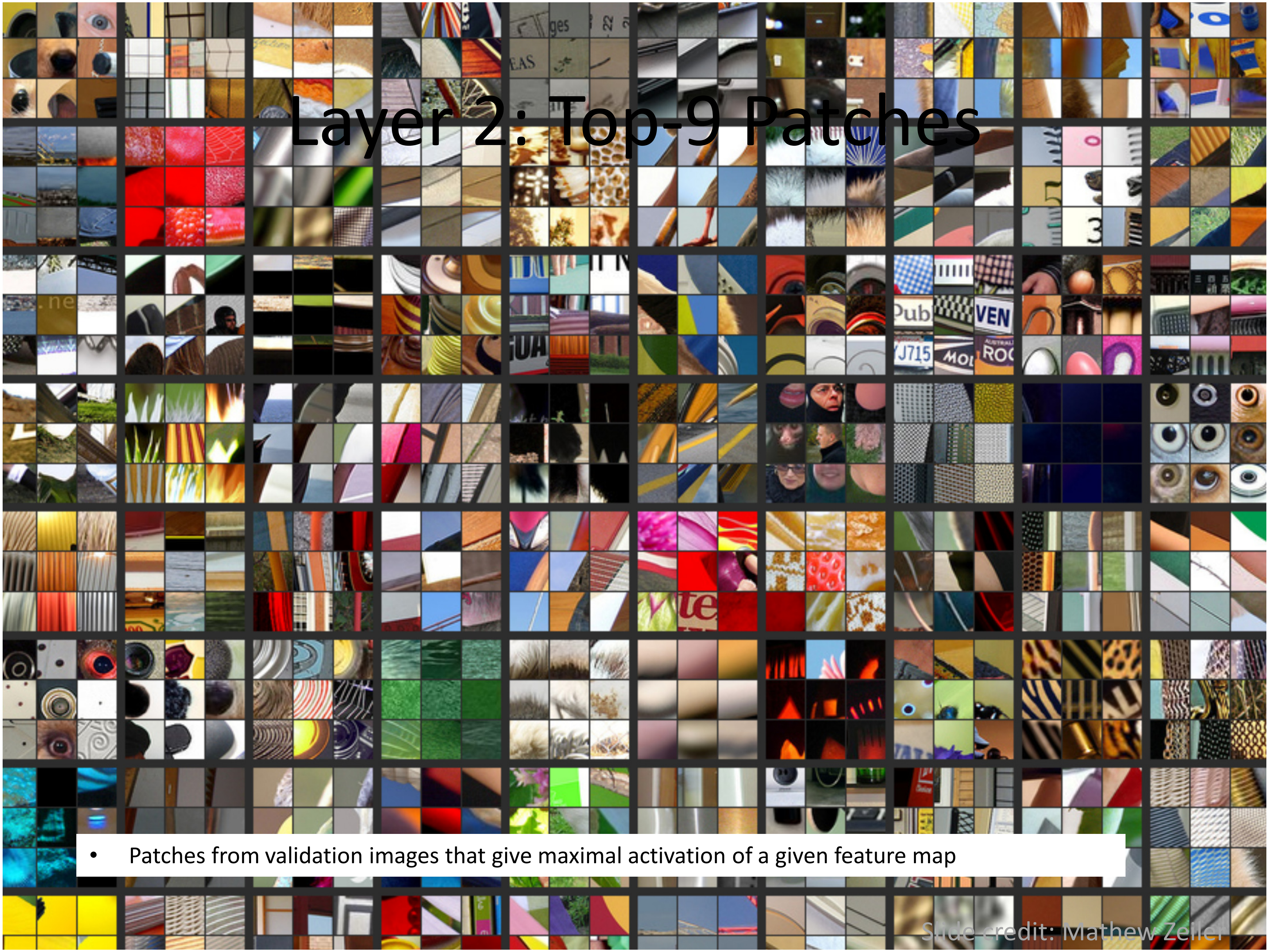
Slide credit: Mathew Zeiler

Layer 2: Top-9 Patches



Slide credit: Mathew Zeiler

Layer 2: Top-9 Patches



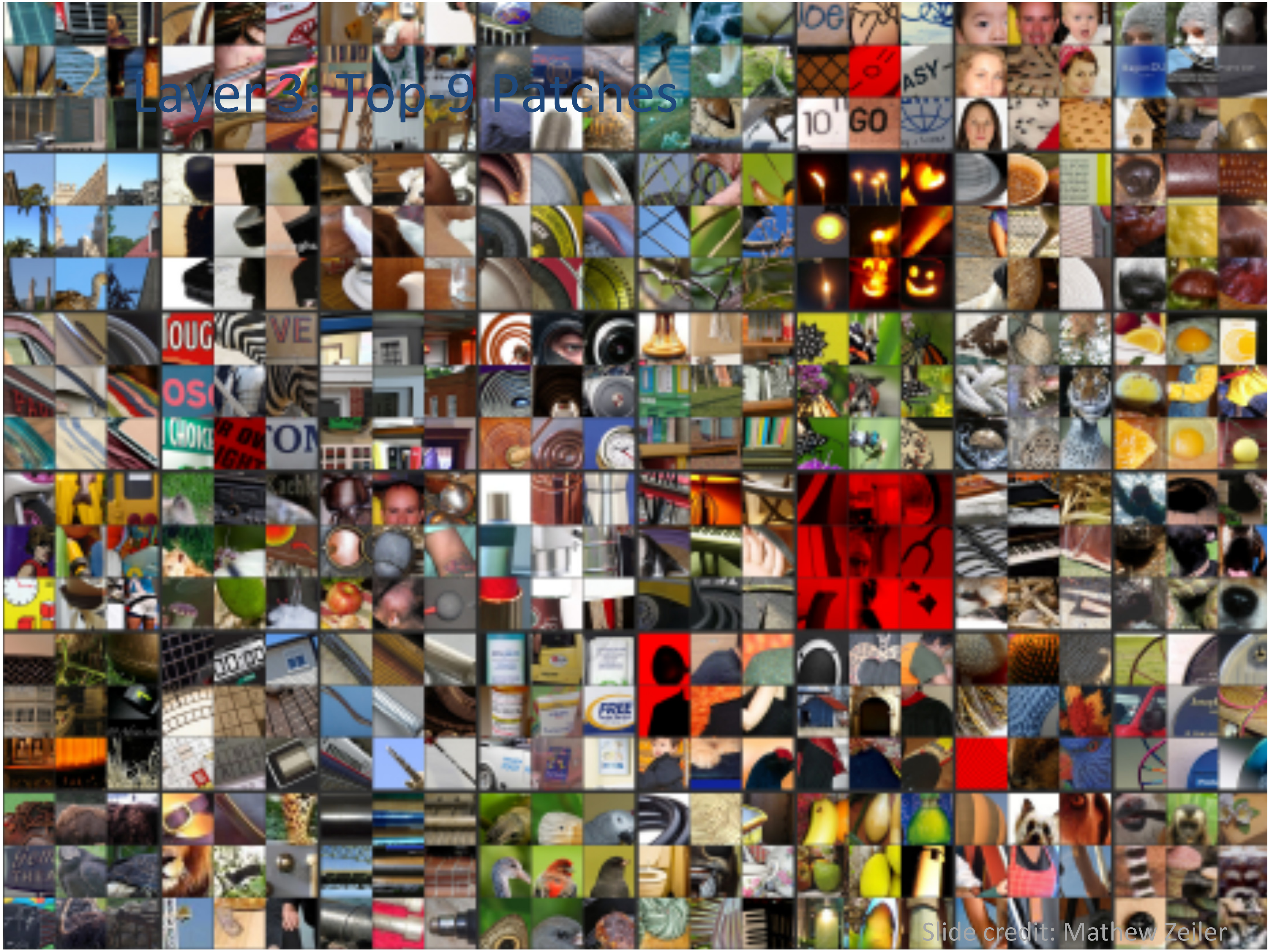
- Patches from validation images that give maximal activation of a given feature map

Slide credit: Mathew Zeiler

Layer 3: Top-9 Patches

Slide credit: Mathew Zeiler

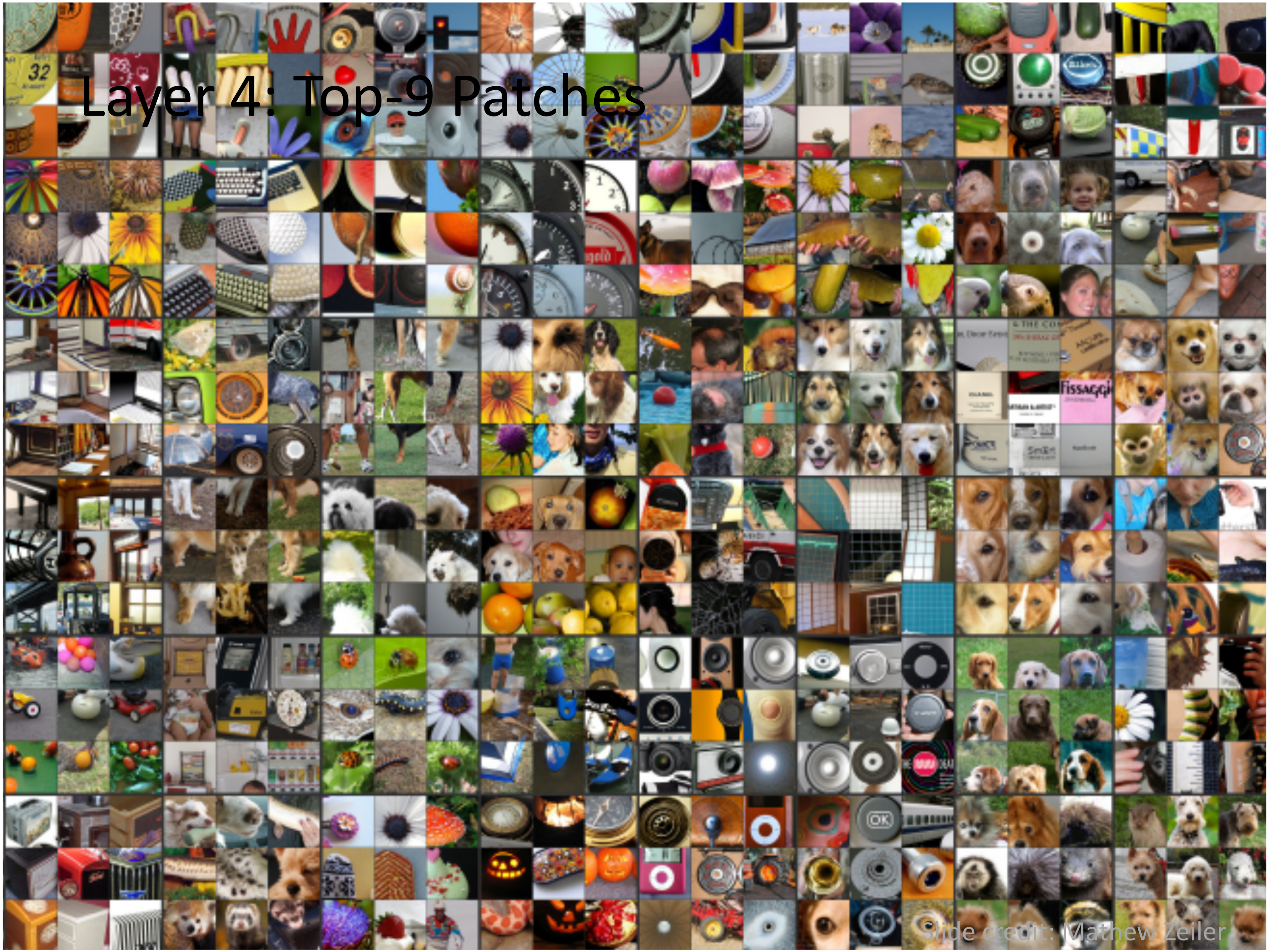
Layer 3: Top-9 Patches



Slide credit: Mathew Zeiler

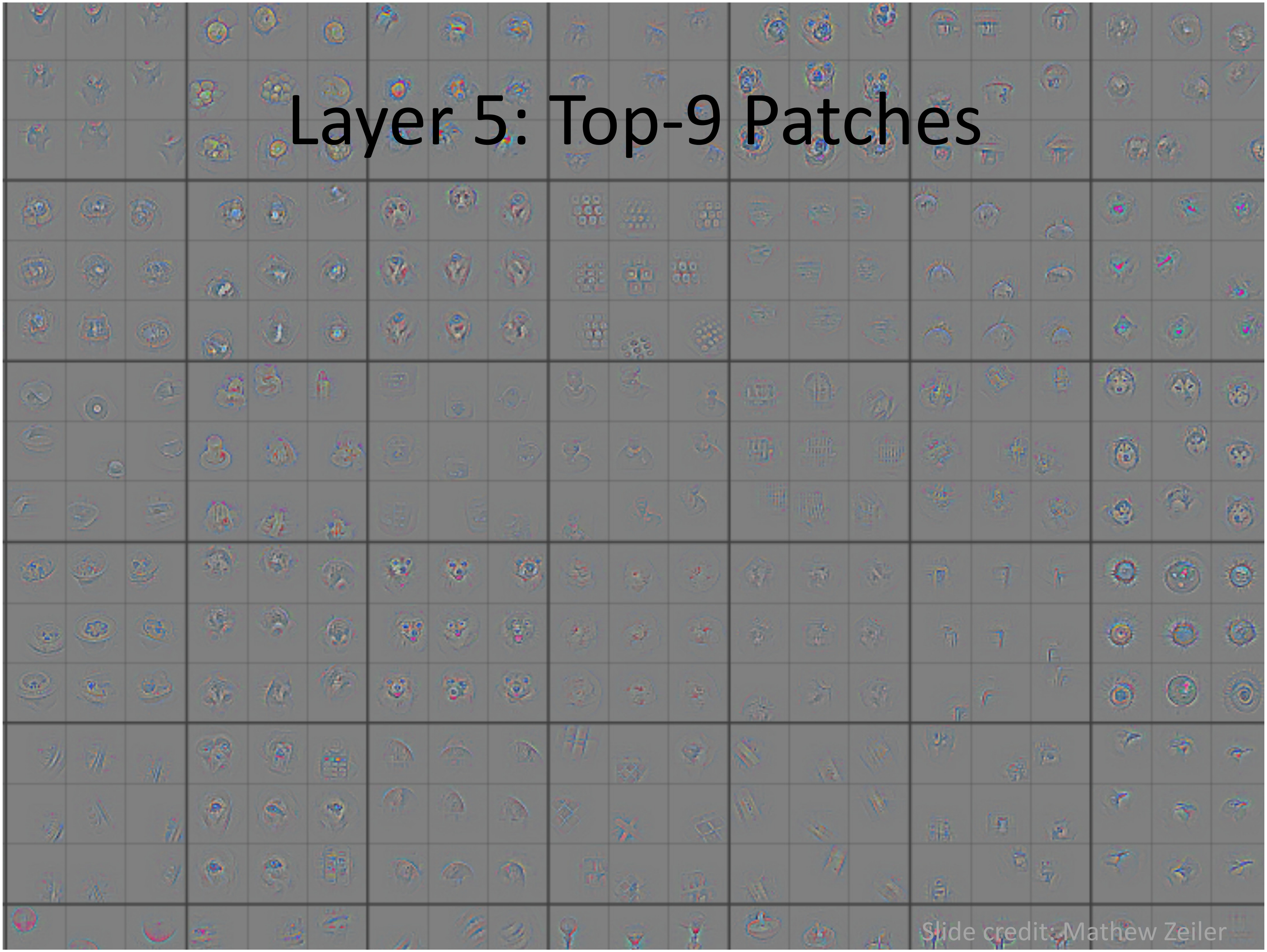
Layer 4: Top-9 Patches

Layer 4: Top-9 Patches



Slide credit: Mathew Zeiler

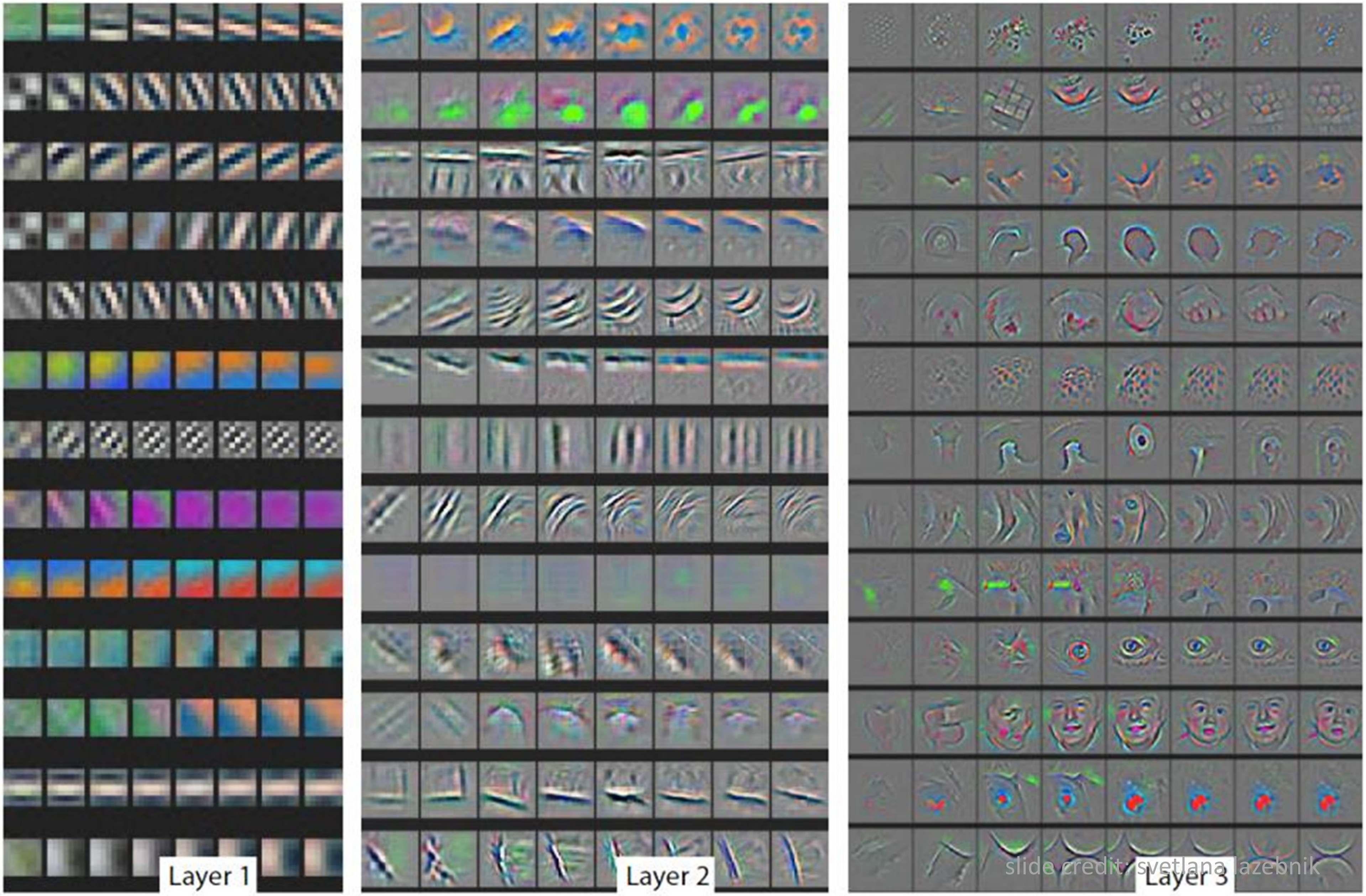
Layer 5: Top-9 Patches



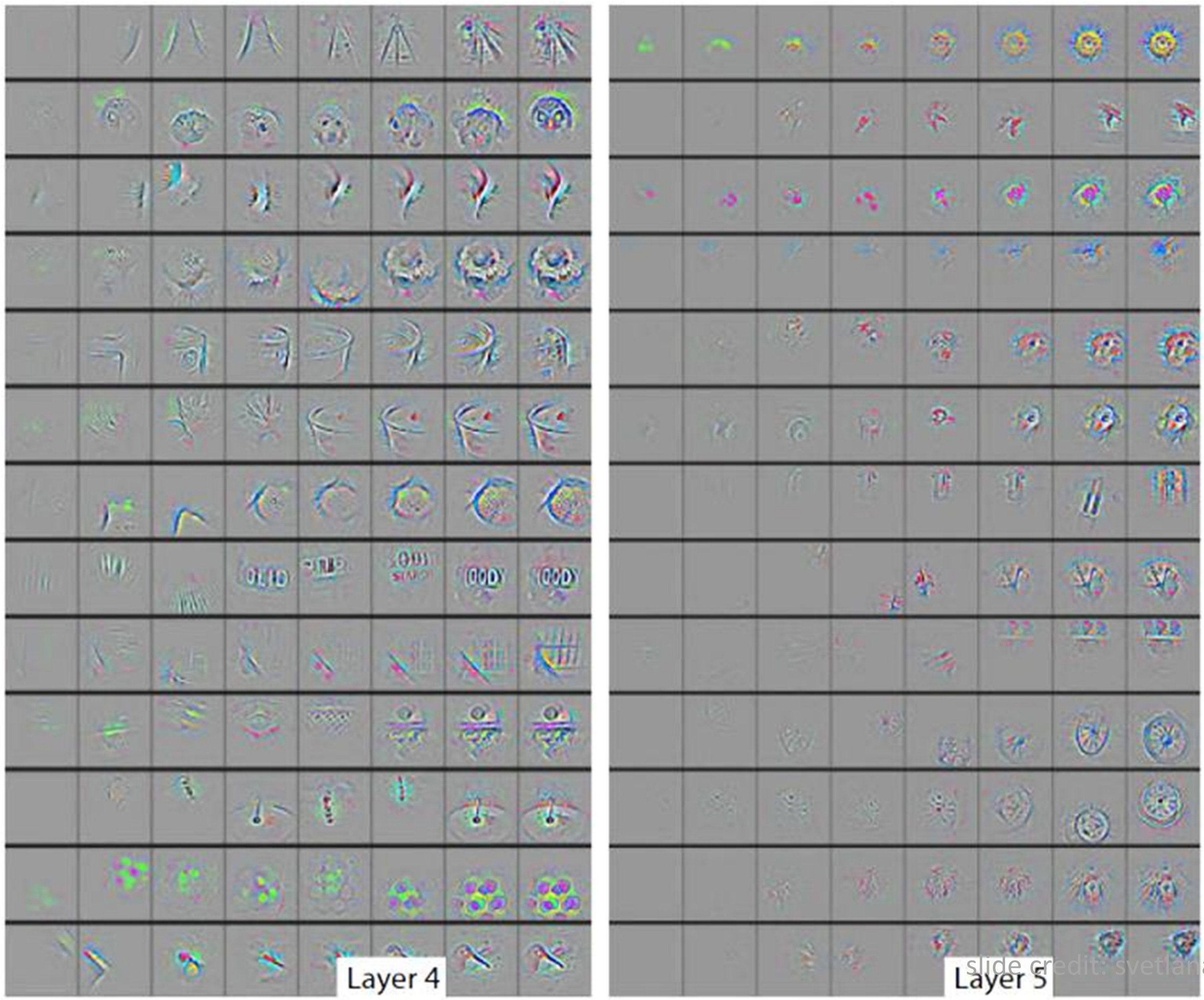
Slide credit: Mathew Zeiler

Layer 5: Top-9 Patches

Evolution of Features During Training

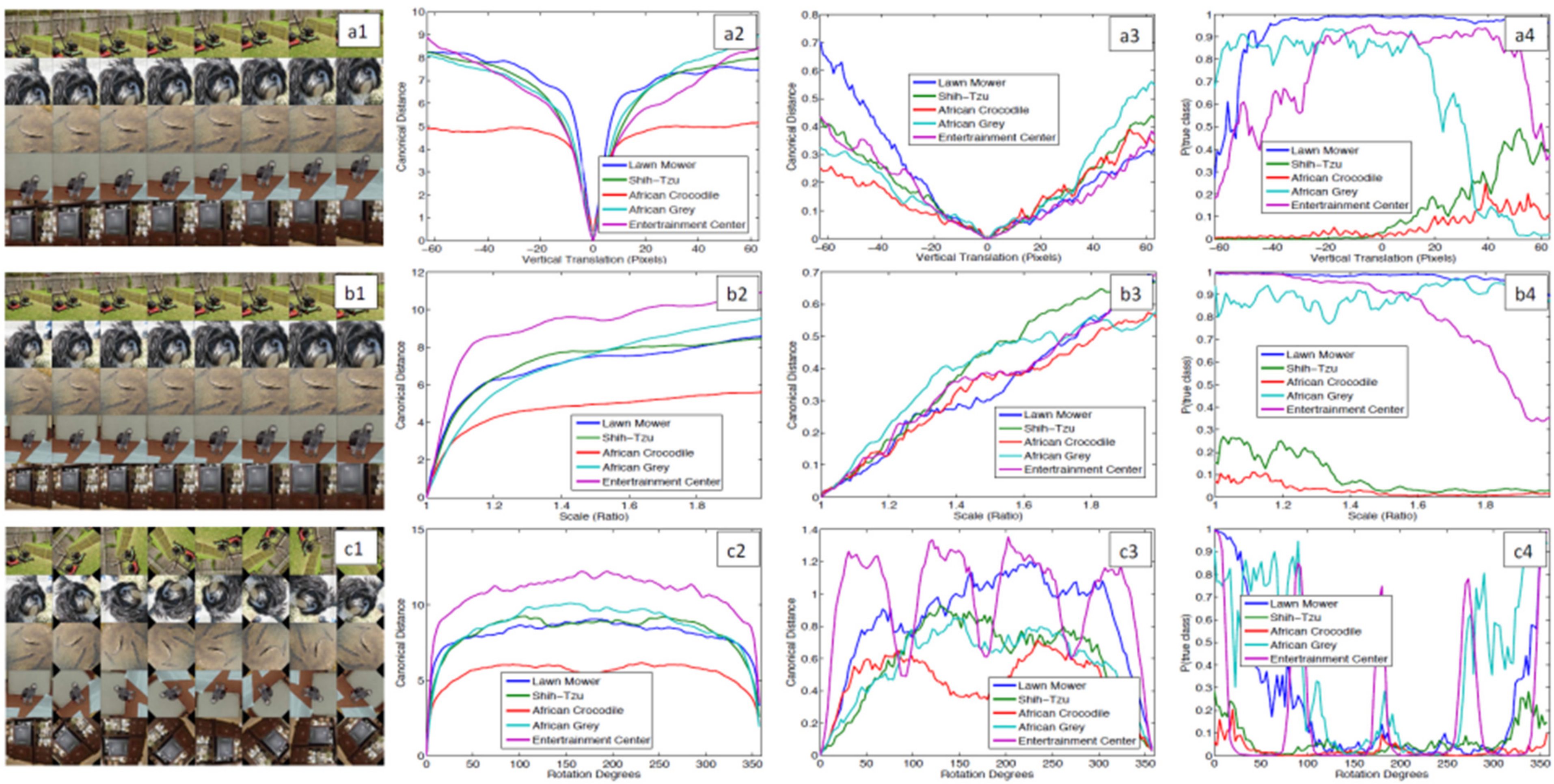


Evolution of Features During Training



slide credit: svetlana lazebnik

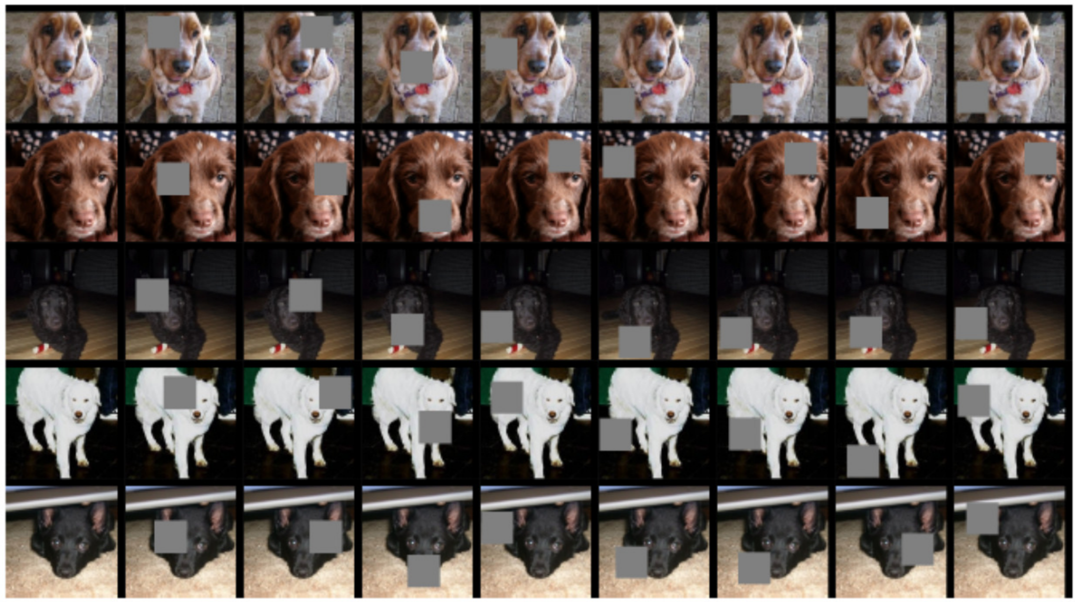
Feature invarience



Layer 1 Layer 7 Prob of true label

slide credit: svetlana lazebnik

Correspondence analysis



Occlusion Location	Mean Feature Sign Change Layer 5	Mean Feature Sign Change Layer 7
Right Eye	0.067 ± 0.007	0.069 ± 0.015
Left Eye	0.069 ± 0.007	0.068 ± 0.013
Nose	0.079 ± 0.017	0.069 ± 0.011
Random	0.107 ± 0.017	0.073 ± 0.014

feature layer
(preserve correspondence)

higher layer
(discriminate different breeds of dog)

slide credit: svetlana lazebnik

Occlusion Experiment

- If the model is truly identifying the location of the object in the image, or just using the surrounding context

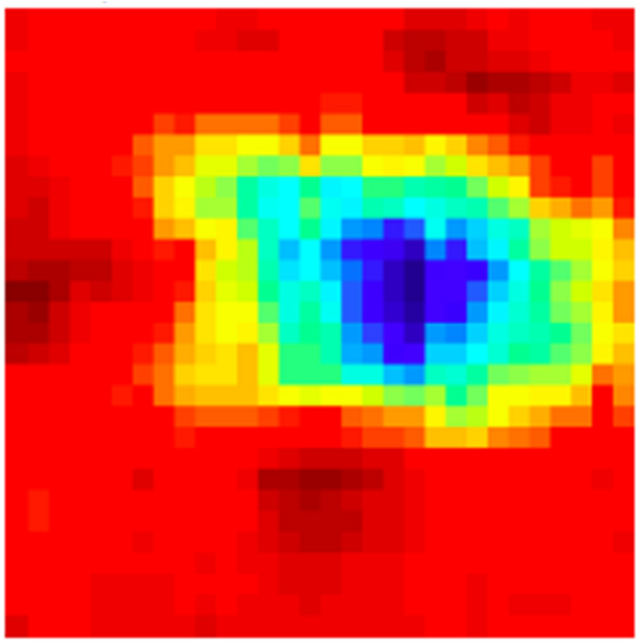
- Mask parts of input with occluding square

- Monitor output

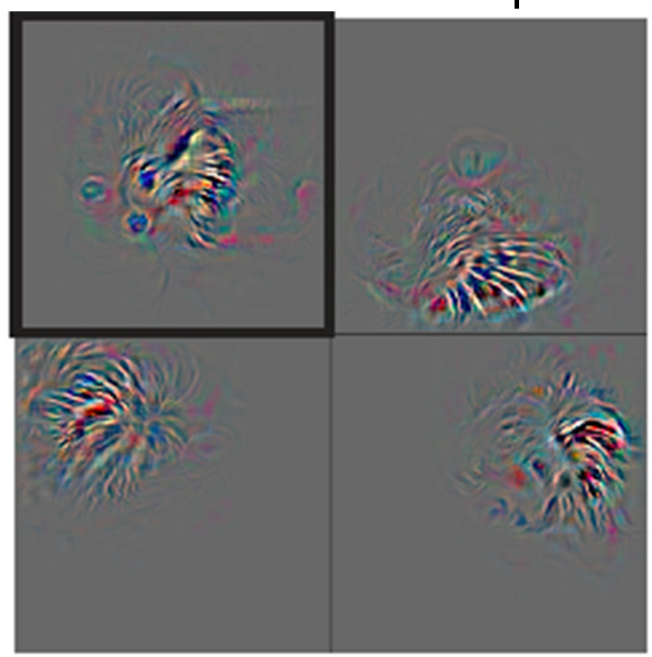
slide credit: svetlana lazebnik

slide credit: svetlana lazebnik

Total activation in most active 5^{th} layer feature map

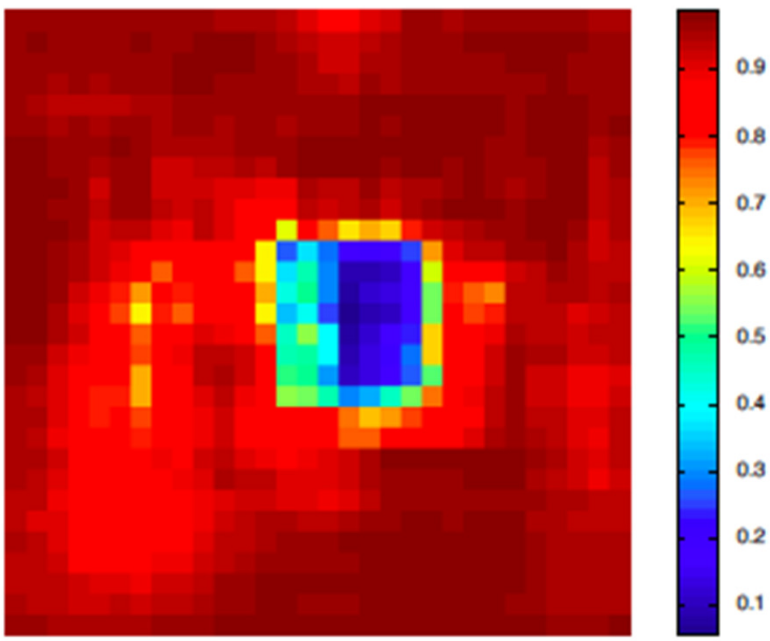


Other activations from same feature map

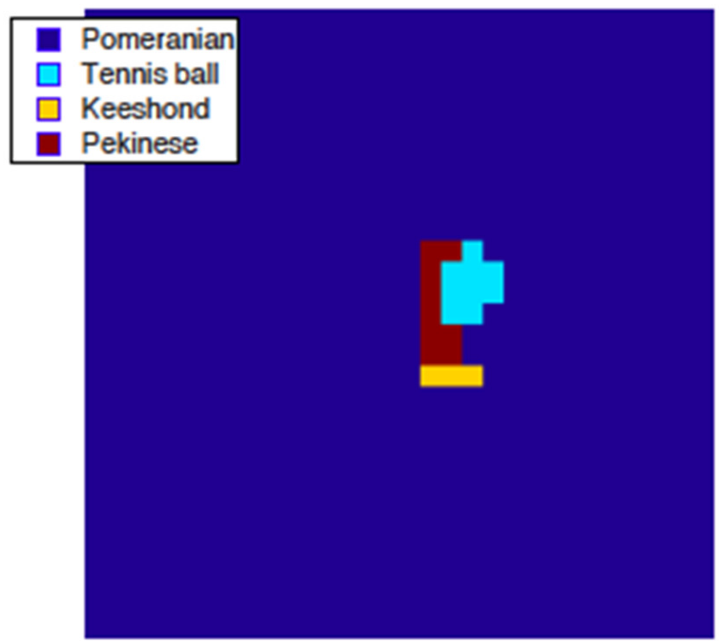


slide credit: svetlana lazebnik

p(True class)

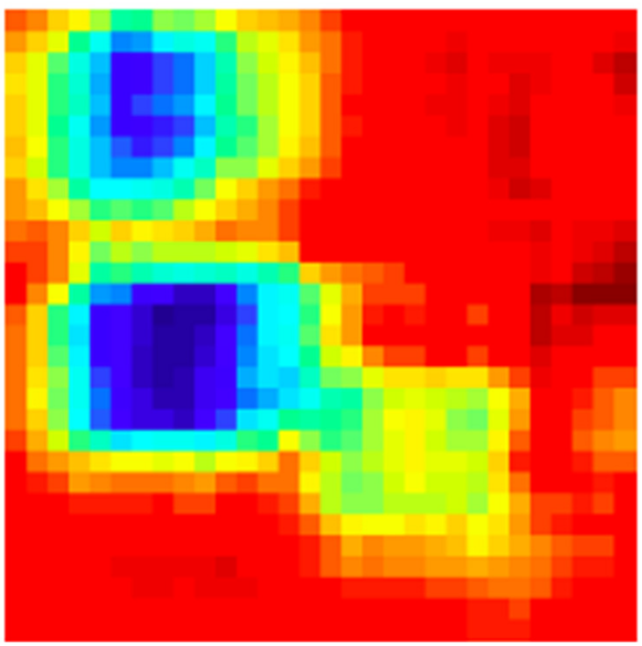


Most probable class



slide credit: svetlana lazebnik

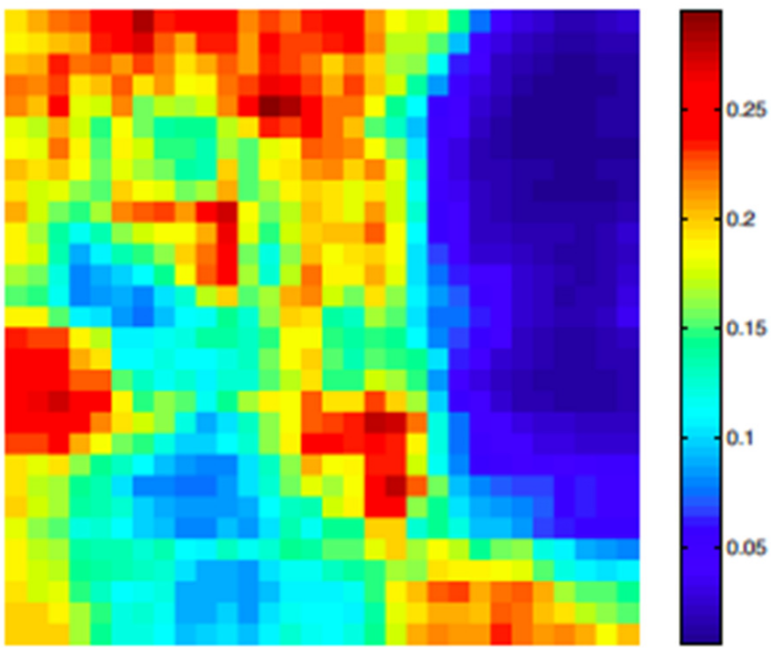
Total activation in most active 5th layer feature map



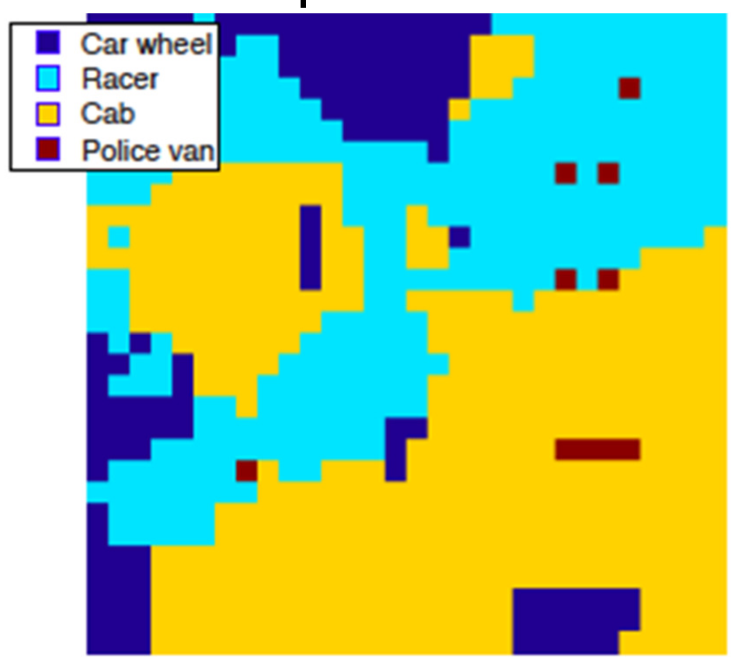
Other activations from same feature map

slide credit: svetlana lazebnik

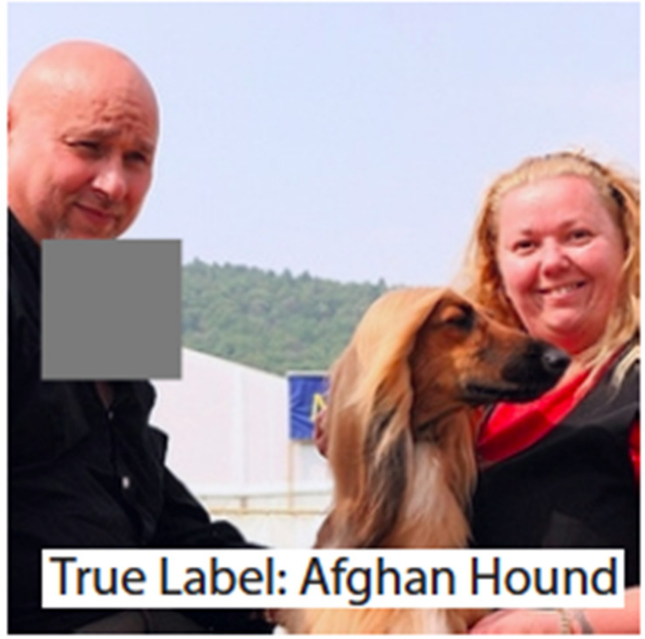
p(True class)



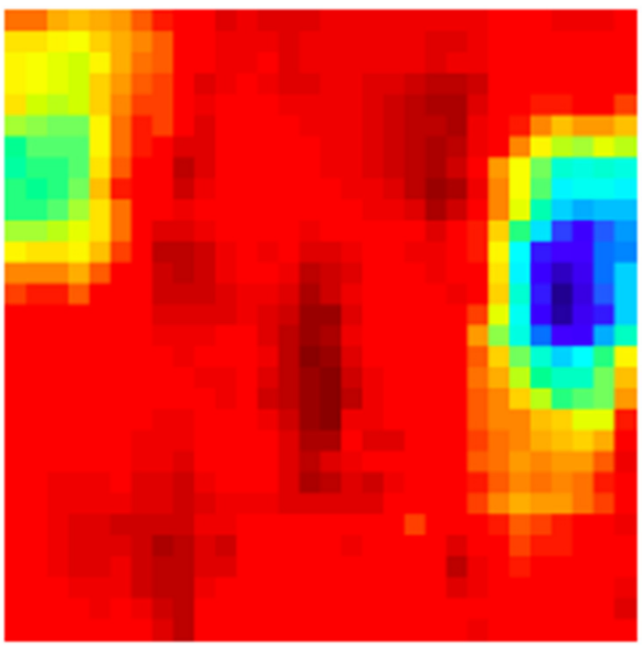
Most probable class



slide credit: svetlana lazebnik

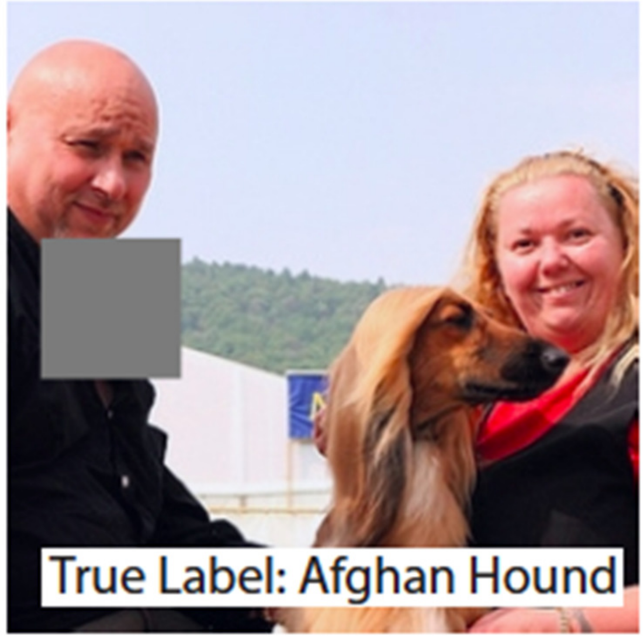


Total activation in most active 5th layer feature map

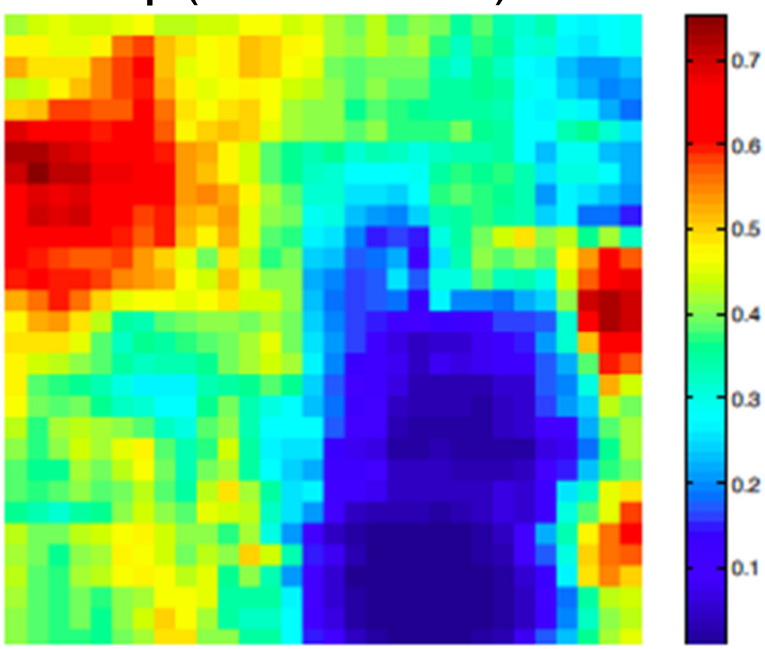


Other activations from same feature map

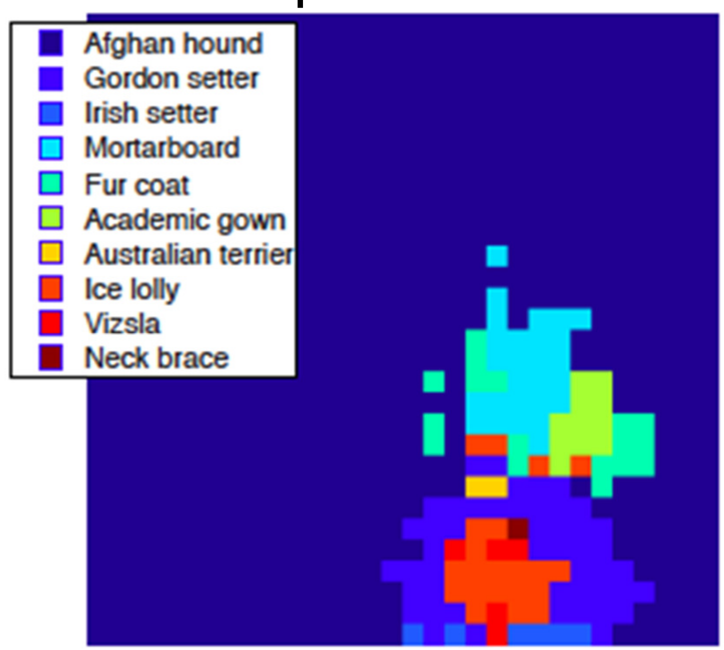
slide credit: svetlana lazebnik

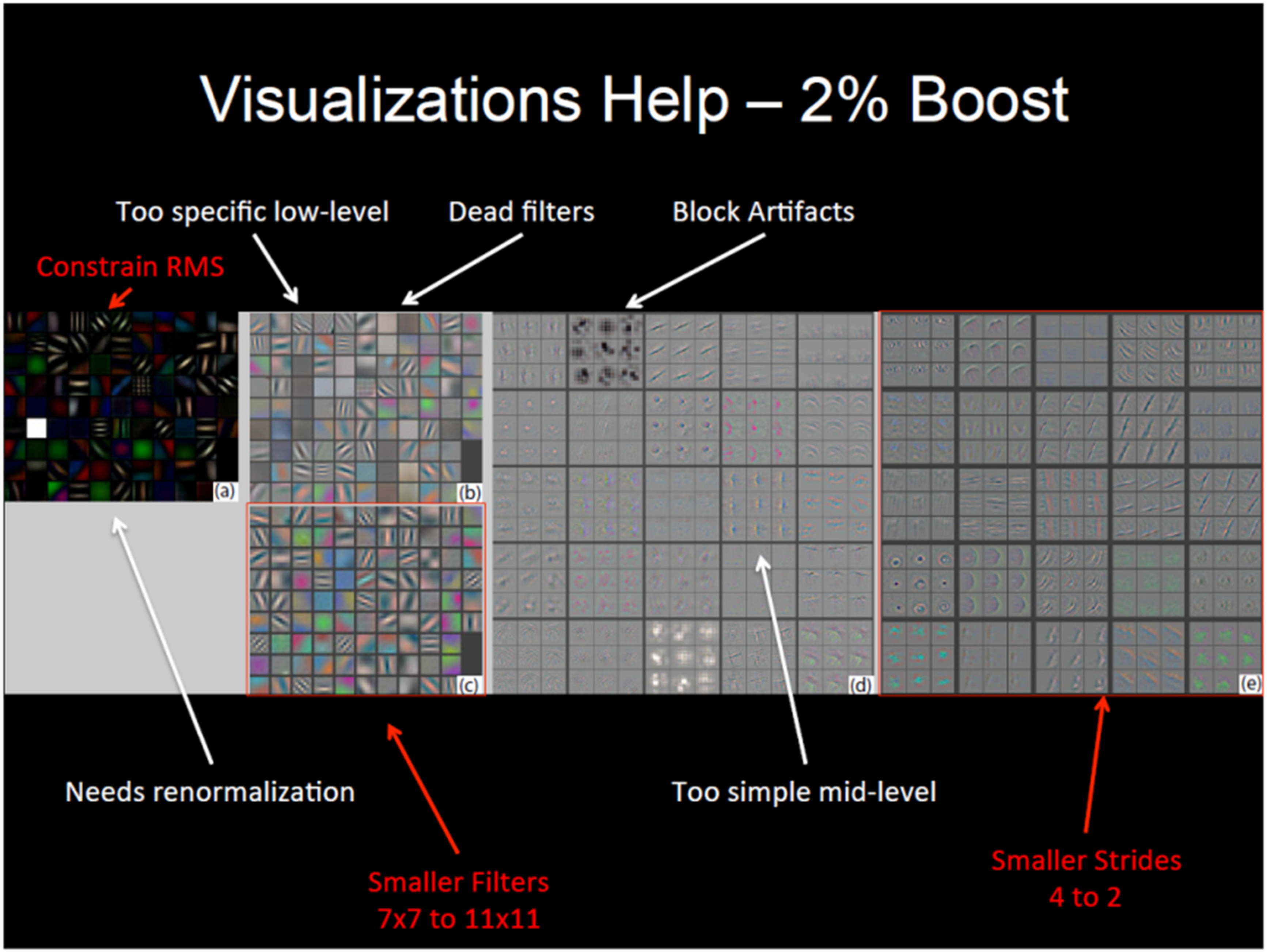


p(True class)



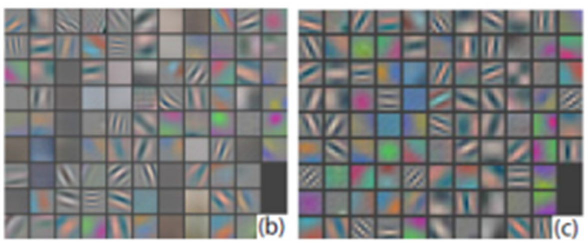
Most probable class



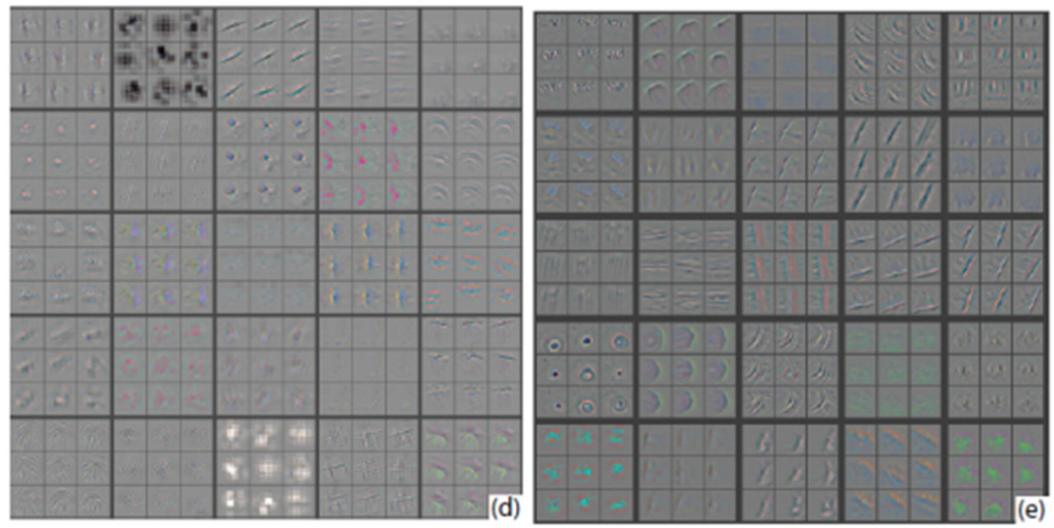


Slide credit: Mathew Zeiler, Rob Fergus

Architecture selection



Layer 1



Layer 2

- Smaller stride (2 vs. 4) and smaller filters (7x7 vs. 11x11)
- Layer 1: more coverage of mid-frequencies
- Layer 2: no aliasing, no “dead” feature

Slide credit: Mathew Zeiler, Rob Fergus

Visualizing Convnets

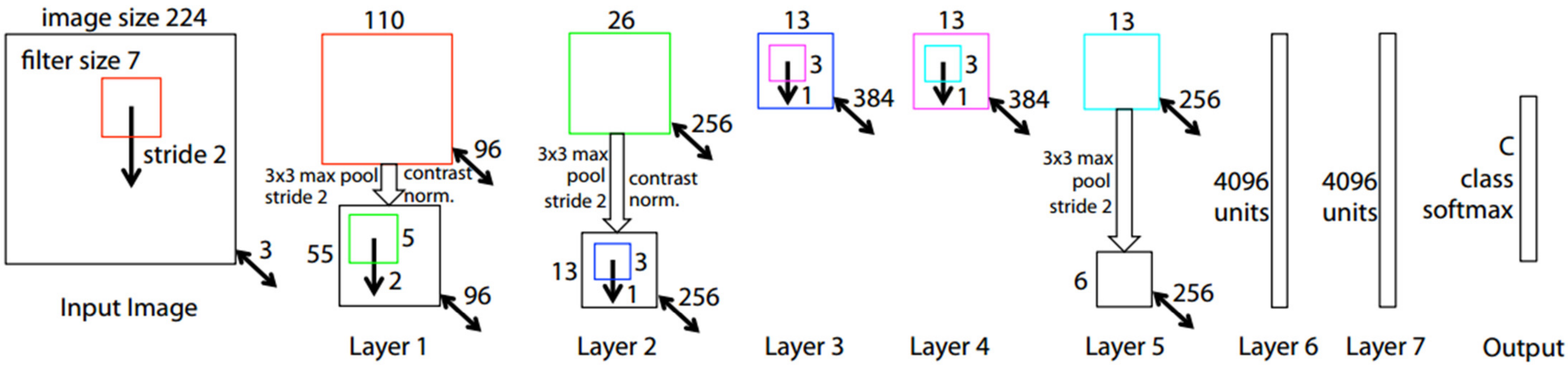


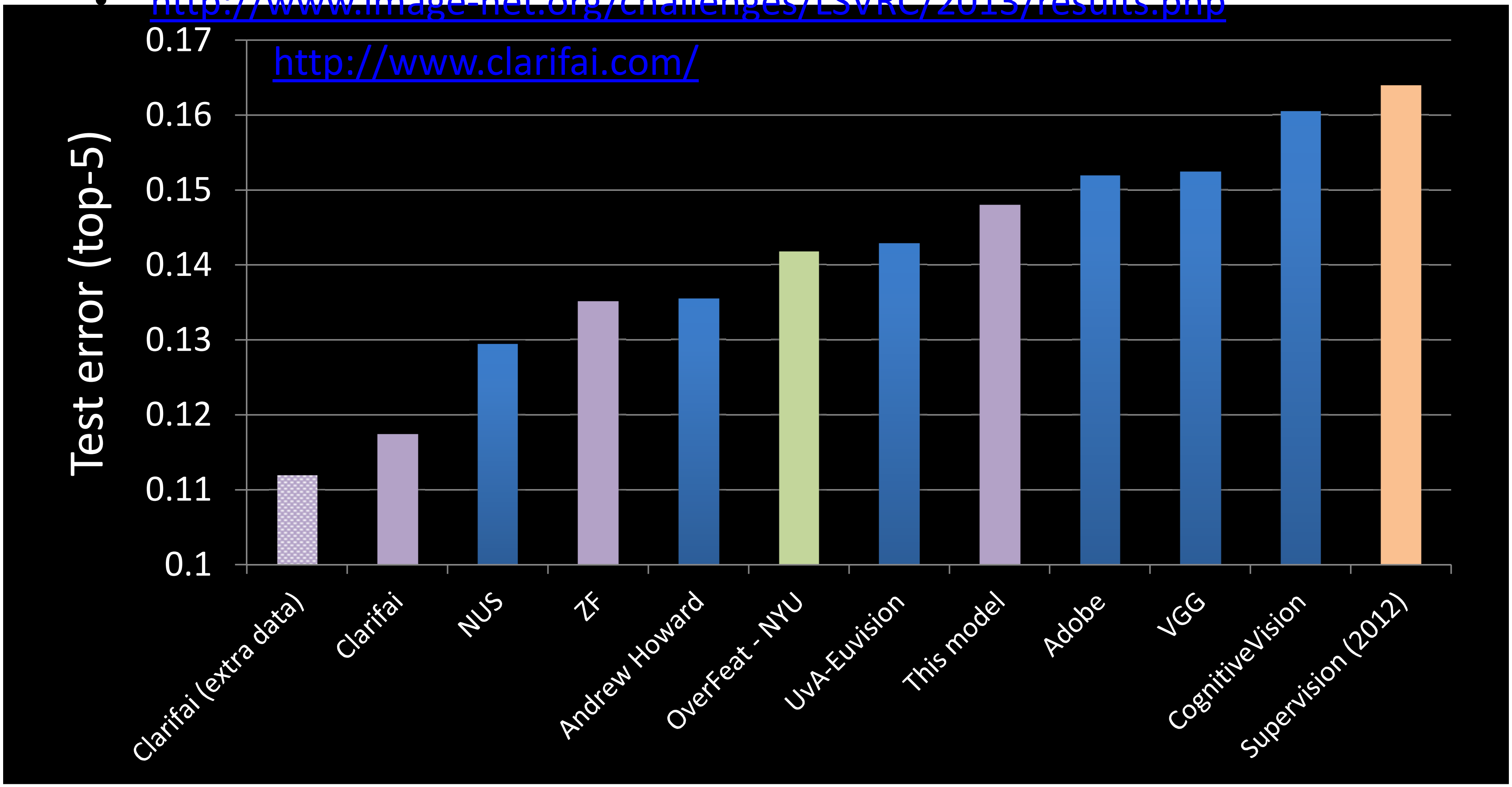
Figure 3. Architecture of our 8 layer convnet model. A 224 by 224 crop of an image (with 3 color planes) is presented as the input. This is convolved with 96 different 1st layer filters (red), each of size 7 by 7, using a stride of 2 in both x and y. The resulting feature maps are then: (i) passed through a rectified linear function (not shown), (ii) pooled (max within 3x3 regions, using stride 2) and (iii) contrast normalized across feature maps to give 96 different 55 by 55 element feature maps. Similar operations are repeated in layers 2,3,4,5. The last two layers are fully connected, taking features from the top convolutional layer as input in vector form ($6 \cdot 6 \cdot 256 = 9216$ dimensions). The final layer is a C-way softmax function, C being the number of classes. All filters and feature maps are square in shape.

M. Zeiler and R. Fergus, [Visualizing and Understanding Convolutional Networks](), arXiv preprint, 2013

Slide credit: Mathew Zeiler, Rob Fergus

ImageNet Classification 2013 Results

- http://www.image-net.org/challenges/LSVRC/2013/results.php
  - http://www.clarifai.com/



Slide credit: Mathew Zeiler

New architecture results

Error %	Val Top-1	Val Top-5	Test Top-5
Gunji *et al.* [12]	-	-	26.2
DeCAF [7]	-	-	19.2
Krizhevsky *et al.* [18], 1 convnet	40.7	18.2	−−
Krizhevsky *et al.* [18], 5 convnets	38.1	16.4	16.4
Krizhevsky *et al.* *[18], 1 convnets	39.0	16.6	−−
Krizhevsky *et al.* *[18], 7 convnets	36.7	15.4	15.3
Our replication of Krizhevsky *et al.* , 1 convnet	40.5	18.1	−−
1 convnet as per Fig. 3	38.4	16.5	−−
5 convnets as per Fig. 3 – (a)	36.7	15.3	15.3
1 convnet as per Fig. 3 but with layers 3,4,5: 512,1024,512 maps – (b)	37.5	16.0	16.1
6 convnets, (a) & (b) combined	36.0	14.7	14.8
Howard [15]	-	-	13.5
Clarifai [28]	-	-	11.7

Classification error rate

Slide credit: Mathew Zeiler, Rob Fergus

Architecture changes

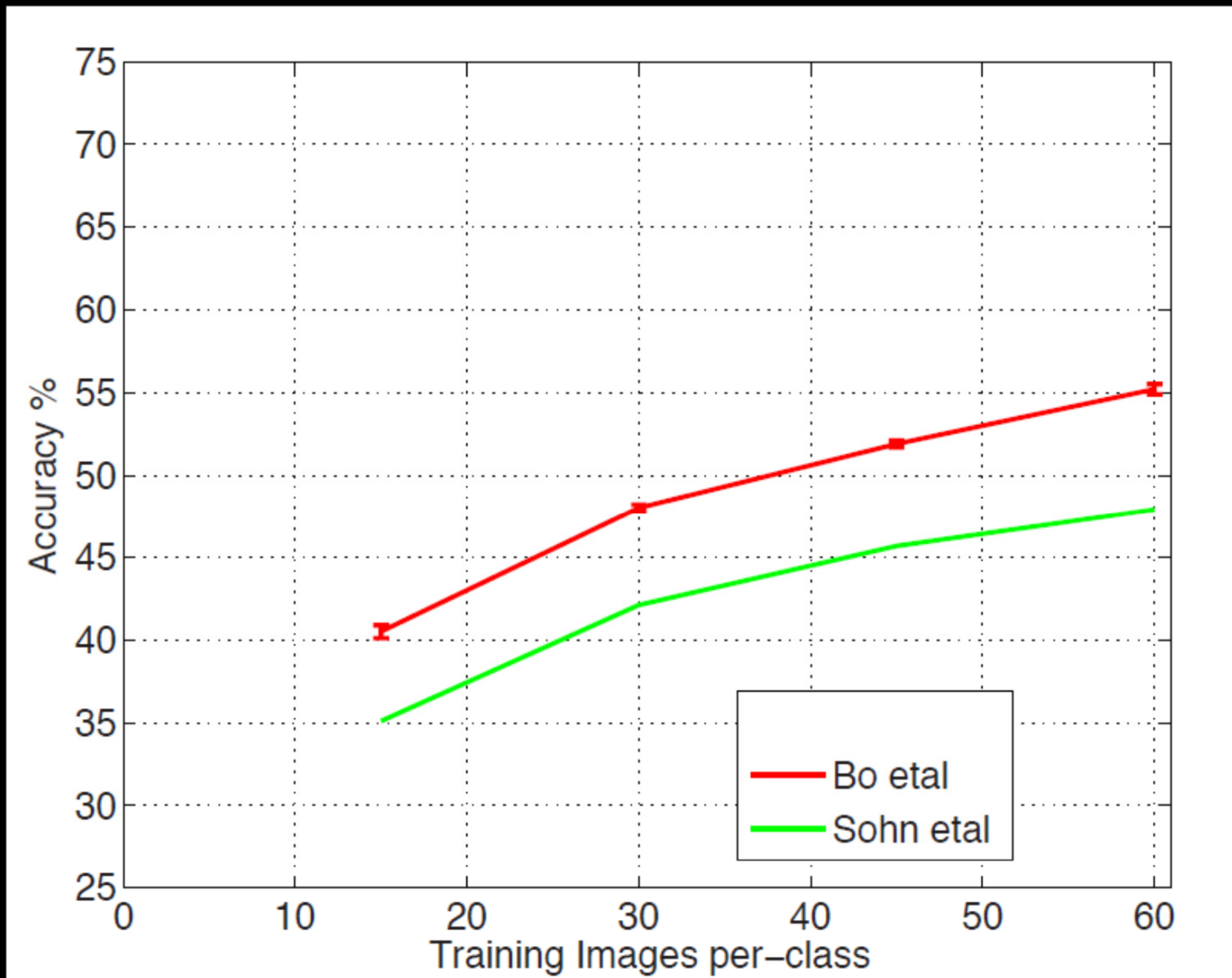
Error %	Train Top-1	Val Top-1	Val Top-5
Our replication of Krizhevsky *et al.* [18], 1 convnet	35.1	40.5	18.1
Removed layers 3,4	41.8	45.4	22.1
Removed layer 7	27.4	40.0	18.4
Removed layers 6,7	27.4	44.8	22.4
Removed layer 3,4,6,7	71.1	71.3	50.1
Adjust layers 6,7: 2048 units	40.3	41.7	18.8
Adjust layers 6,7: 8192 units	26.8	40.0	18.1
Our Model (as per Fig. 3)	33.1	38.4	16.5
Adjust layers 6,7: 2048 units	38.2	40.2	17.6
Adjust layers 6,7: 8192 units	22.0	38.8	17.0
Adjust layers 3,4,5: 512,1024,512 maps	18.8	**37.5**	**16.0**
Adjust layers 6,7: 8192 units and Layers 3,4,5: 512,1024,512 maps	**10.0**	38.3	16.9

increase size of convolution layers

Classification error rate

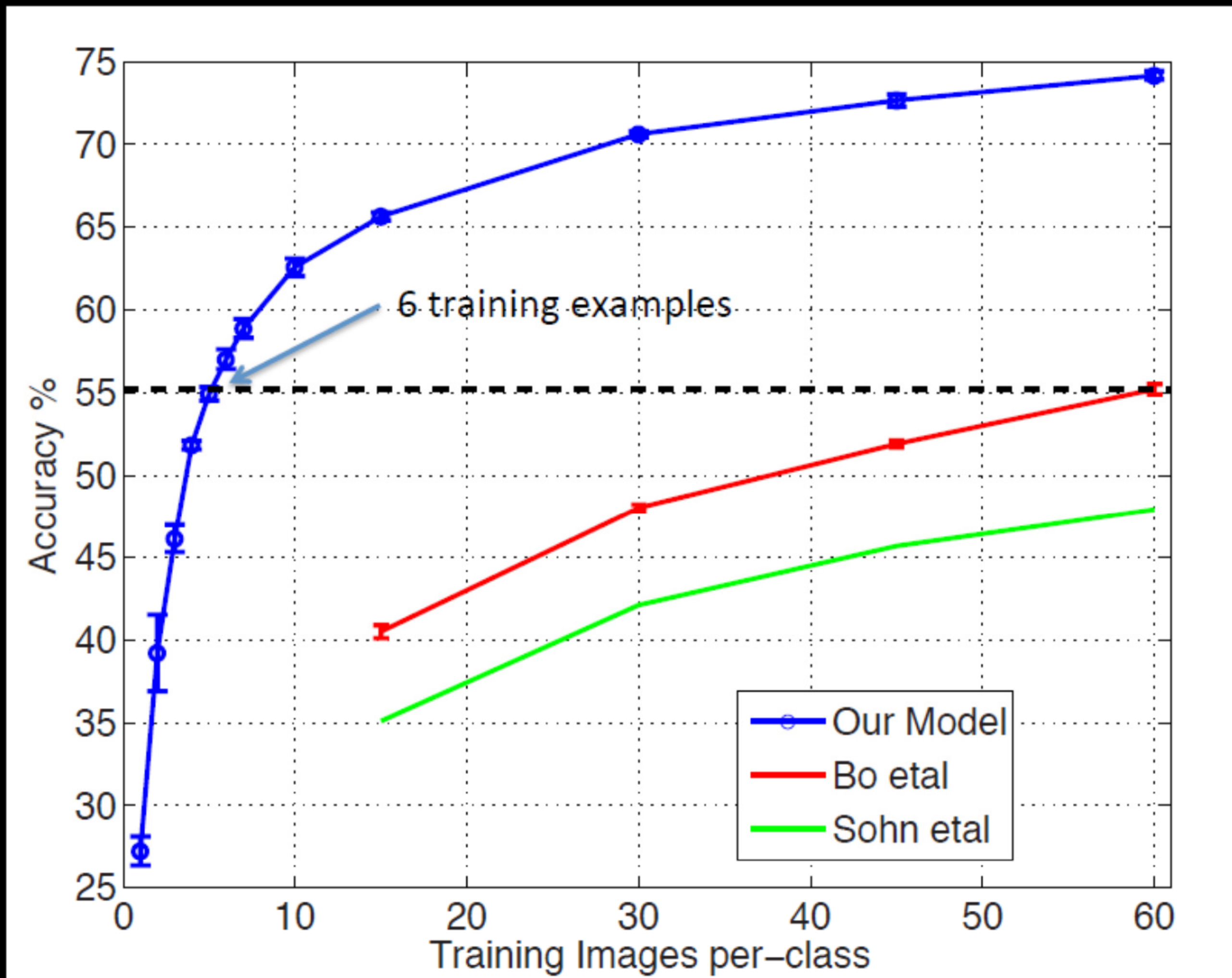
Slide credit: Mathew Zeiler, Rob Fergus

Caltech 256



Slide credit: Mathew Zeiler

Caltech 256



Slide credit: Mathew Zeiler

Results

# Train	Acc % 15/class	Acc % 30/class
(Bo et al., 2013)	–	81.4 ± 0.33
(Jianchao et al., 2009)	73.2	84.3
Non-pretrained convnet	22.8 ± 1.5	46.5 ± 1.7
ImageNet-pretrained convnet	$\mathbf{83.8 \pm 0.5}$	$\mathbf{86.5 \pm 0.5}$

Caltech 101

# Train	Acc % 15/class	Acc % 30/class	Acc % 45/class	Acc % 60/class
(Sohn et al., 2011)	35.1	42.1	45.7	47.9
(Bo et al., 2013)	40.5 ± 0.4	48.0 ± 0.2	51.9 ± 0.2	55.2 ± 0.3
Non-pretr.	9.0 ± 1.4	22.5 ± 0.7	31.2 ± 0.5	38.8 ± 1.4
ImageNet-pretr.	$\mathbf{65.7 \pm 0.2}$	$\mathbf{70.6 \pm 0.2}$	$\mathbf{72.7 \pm 0.4}$	$\mathbf{74.2 \pm 0.3}$

Caltech 256

Results

Acc %	[A]	[B]	Ours	Acc %	[A]	[B]	Ours
Airplane	92.0	**97.3**	96.0	Dining tab	63.2	**77.8**	67.7
Bicycle	74.2	**84.2**	77.1	Dog	68.9	83.0	**87.8**
Bird	73.0	80.8	**88.4**	Horse	78.2	**87.5**	86.0
Boat	77.5	85.3	**85.5**	Motorbike	81.0	**90.1**	85.1
Bottle	54.3	**60.8**	55.8	Person	91.6	**95.0**	90.9
Bus	85.2	**89.9**	85.8	Potted pl	55.9	**57.8**	52.2
Car	81.9	**86.8**	78.6	Sheep	69.4	79.2	**83.6**
Cat	76.4	89.3	**91.2**	Sofa	65.4	**73.4**	61.1
Chair	65.2	**75.4**	65.0	Train	86.7	**94.5**	91.8
Cow	63.2	**77.8**	74.4	Tv	77.4	**80.7**	76.1
Mean	74.3	**82.2**	79.0	# won	0	**15**	5

Pascal VOC

([A]= (Sande et al., 2012) and [B] = (Yan et al., 2012))

- Acknowledgement
  - Slides taken from
    - Matt Zieler
    - Rob Fergus
    - Svetlana Lazebnik