Visualizing and Understanding Convolutional Networks

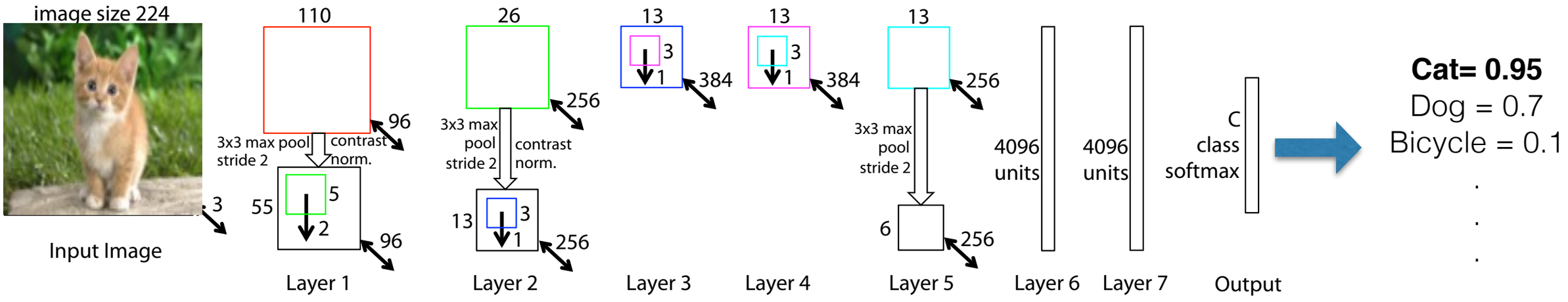
Authors: Matthew D. Zeiler and Rob Fergus
New York University

Presenter: Hamid Izadinia

Vision seminar (Autumn 2014)

Contributions

- Impressive classification performance of CNN
- No clear understanding why
- Introduce network activation visualization
- Diagnostic the effect of each layer & setting
- Find an optimum architecture
- Occlusion experiments for spatial understanding



Hierarchical Convolution, Nonlinear operations (ReLU, max pooling)

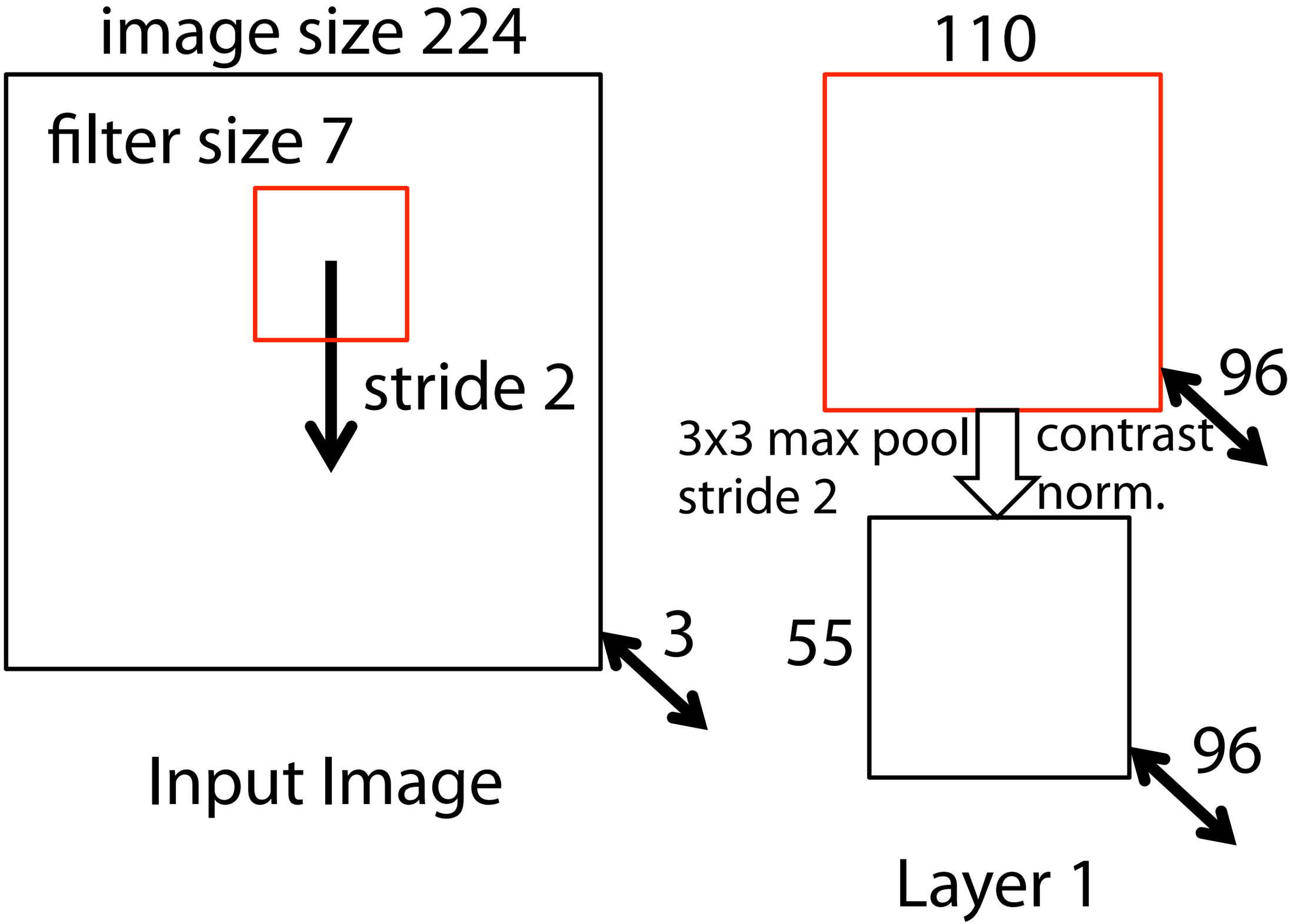


image size 224
filter size 7
stride 2
3
Input Image
110
96
3x3 max pool
stride 2
contrast
norm.
55
96
Layer 1

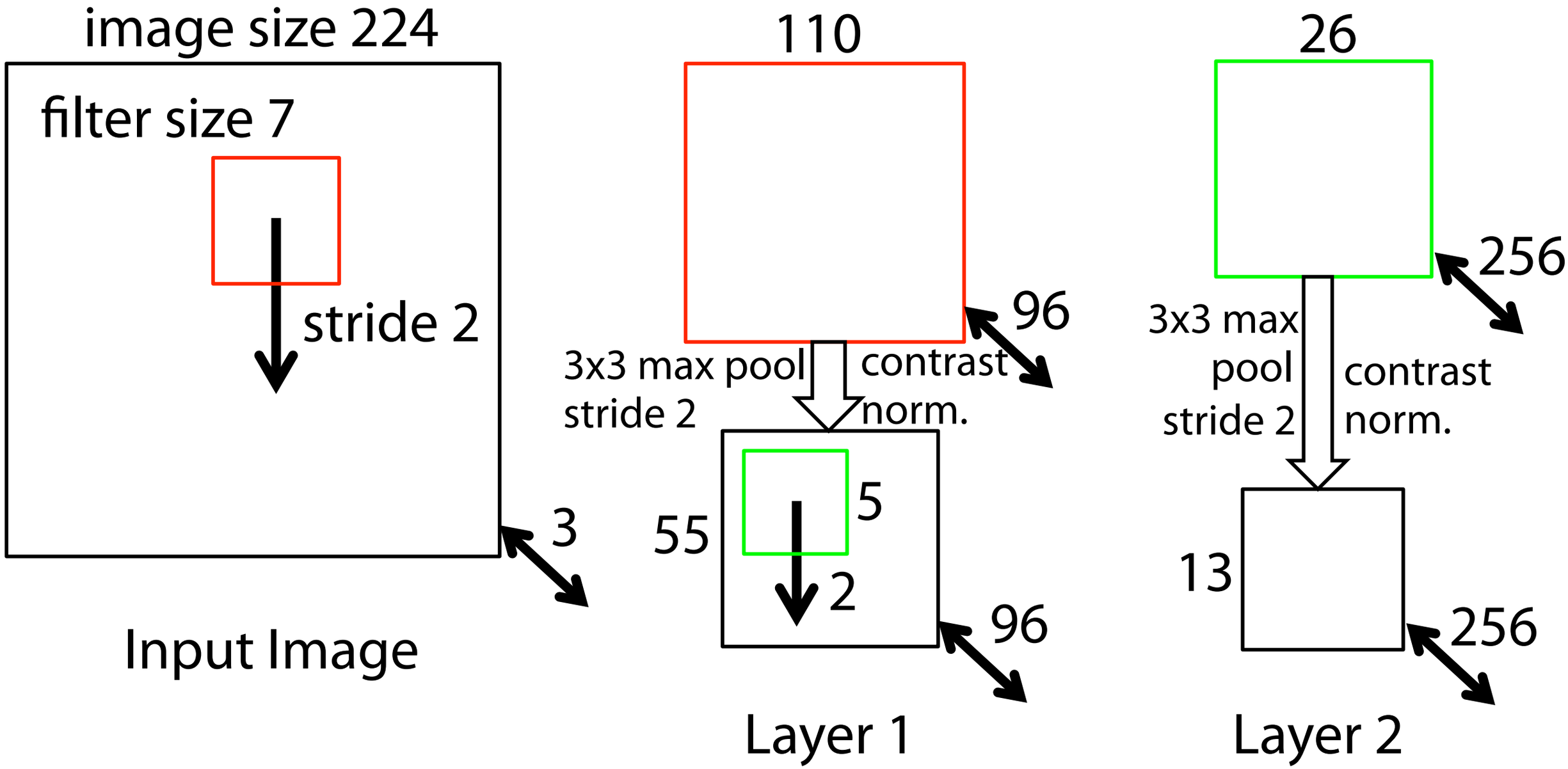
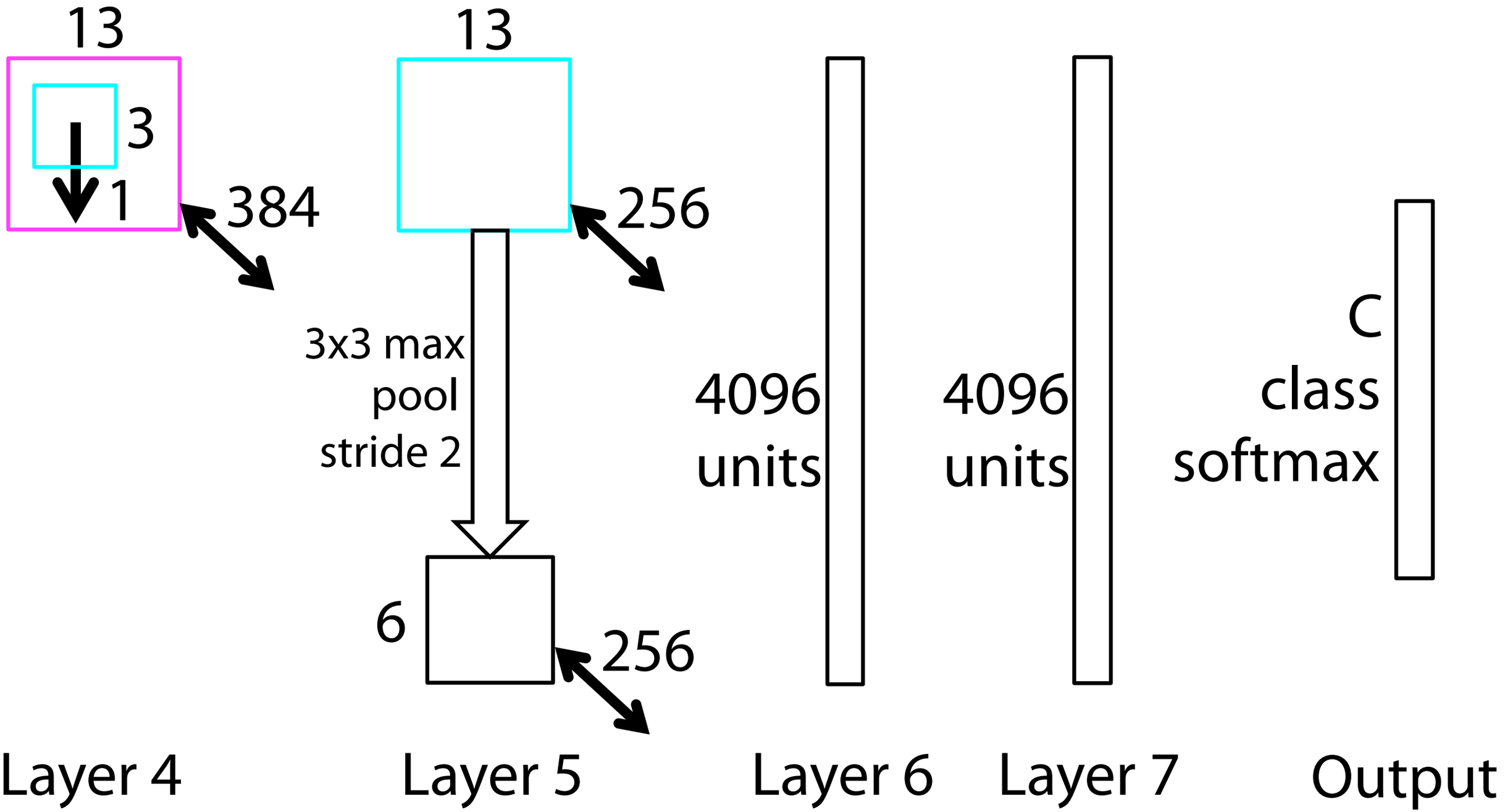
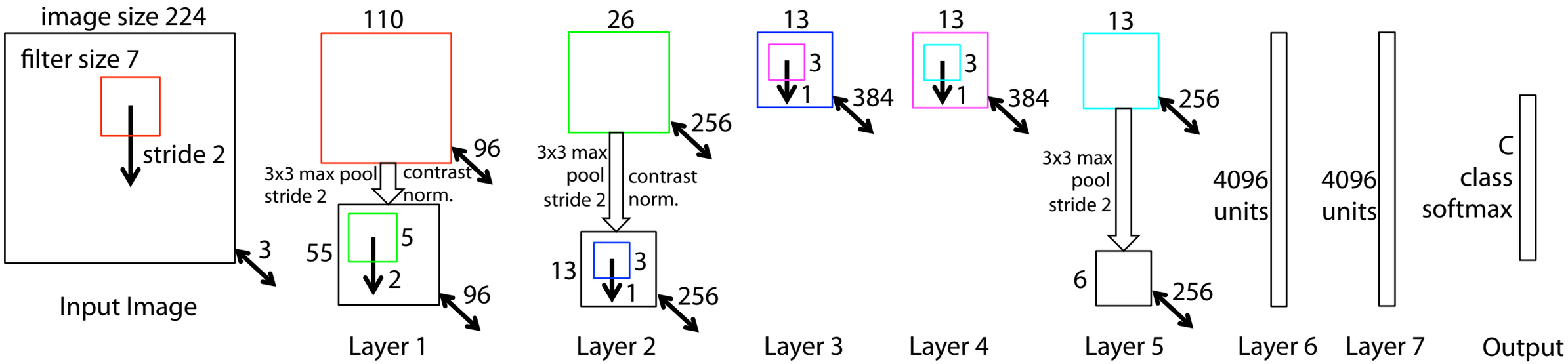


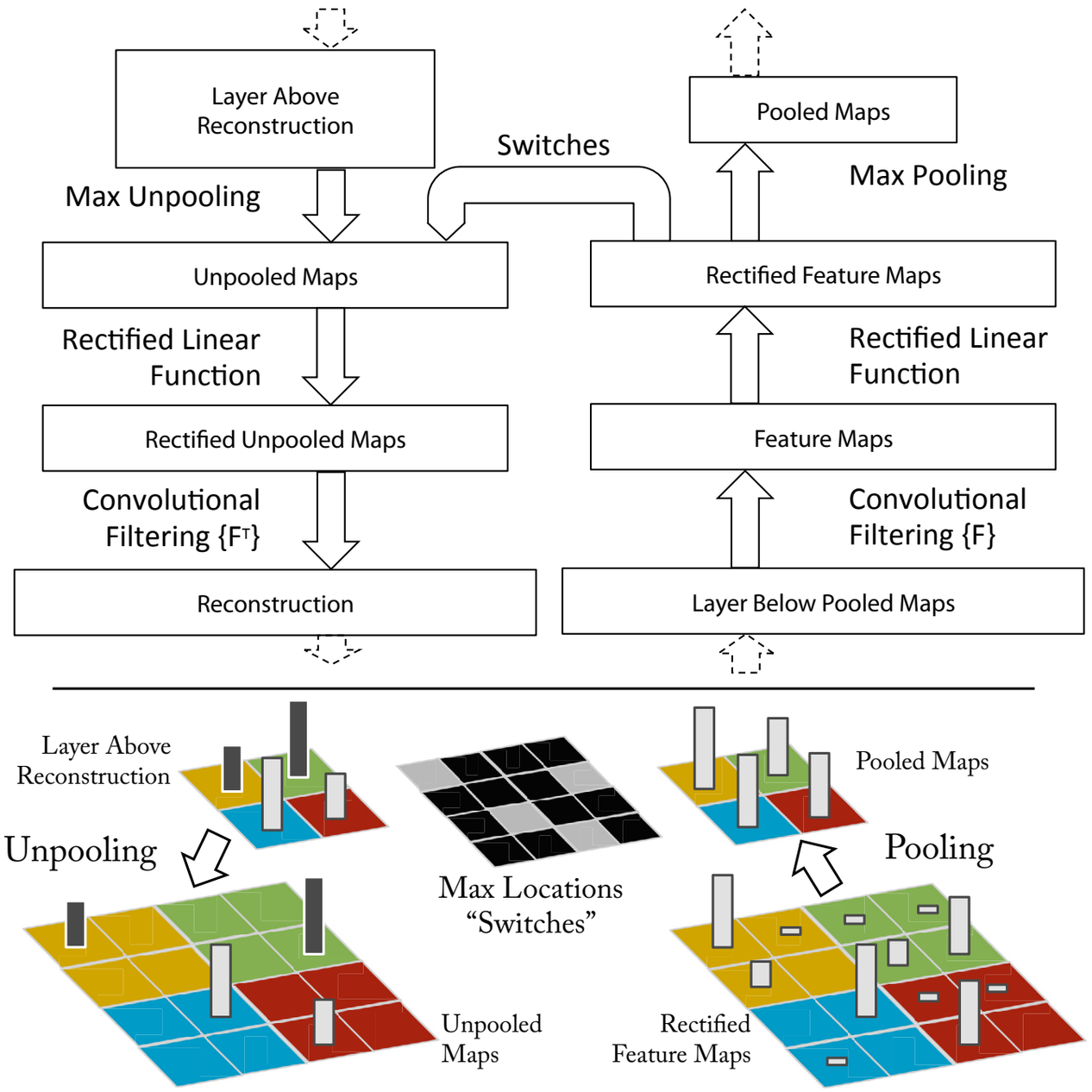
image size 224
filter size 7
stride 2
3
Input Image
110
3x3 max pool
stride 2
contrast
norm.
96
55
5
2
96
Layer 1
26
3x3 max
pool
stride 2
contrast
norm.
256
13
256
Layer 2


13
13
3
1
384
256
3x3 max
pool
stride 2
6
256
4096
units
4096
units
C
class
softmax
Layer 4
Layer 5
Layer 6
Layer 7
Output

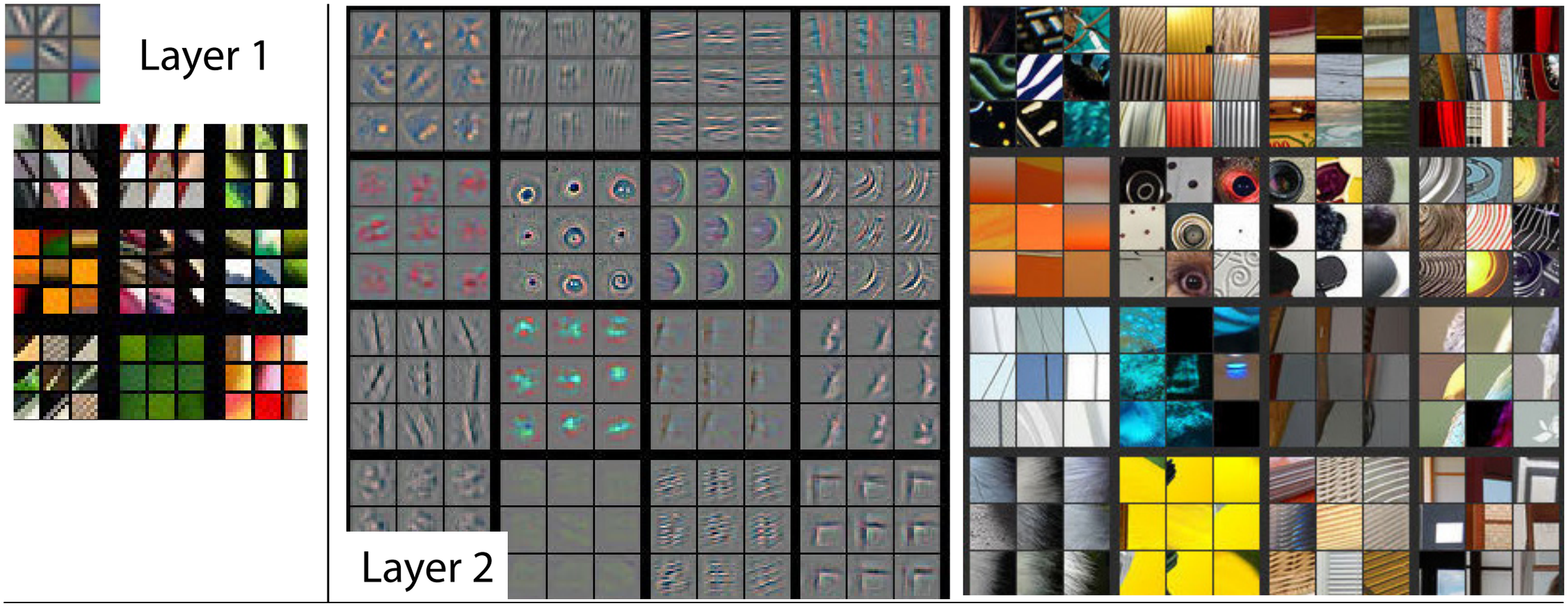
Convolutional Neural Network



Deconvnet & Convnet

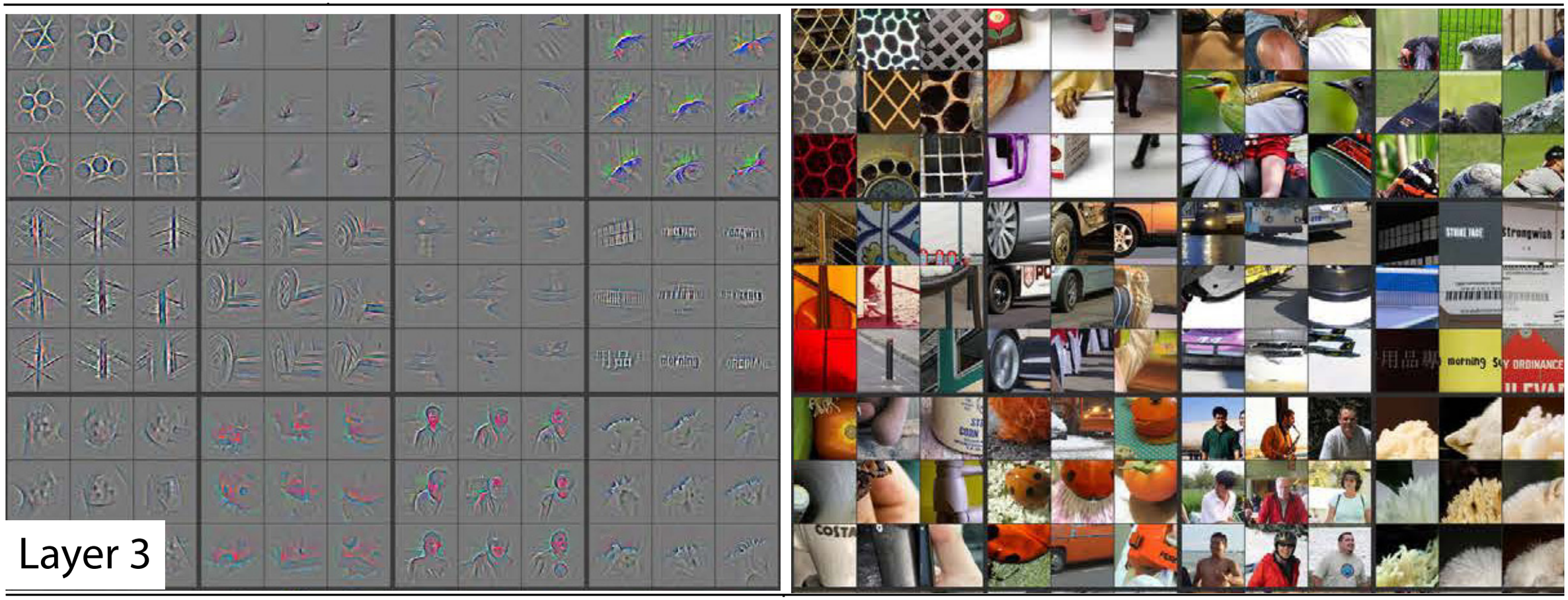


Feature visualization



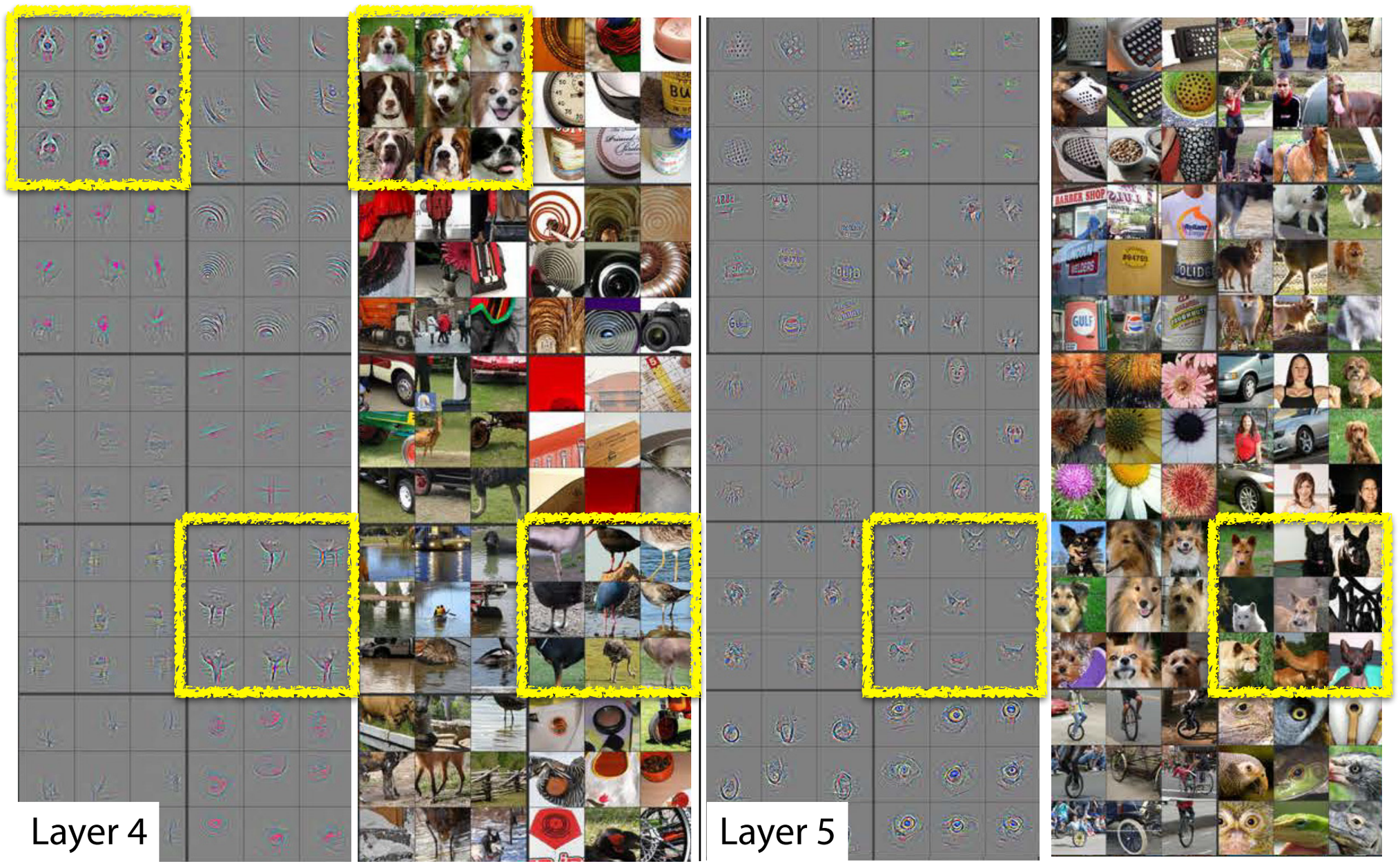
corners & edge/color conjunctions

Feature visualization



similar textures

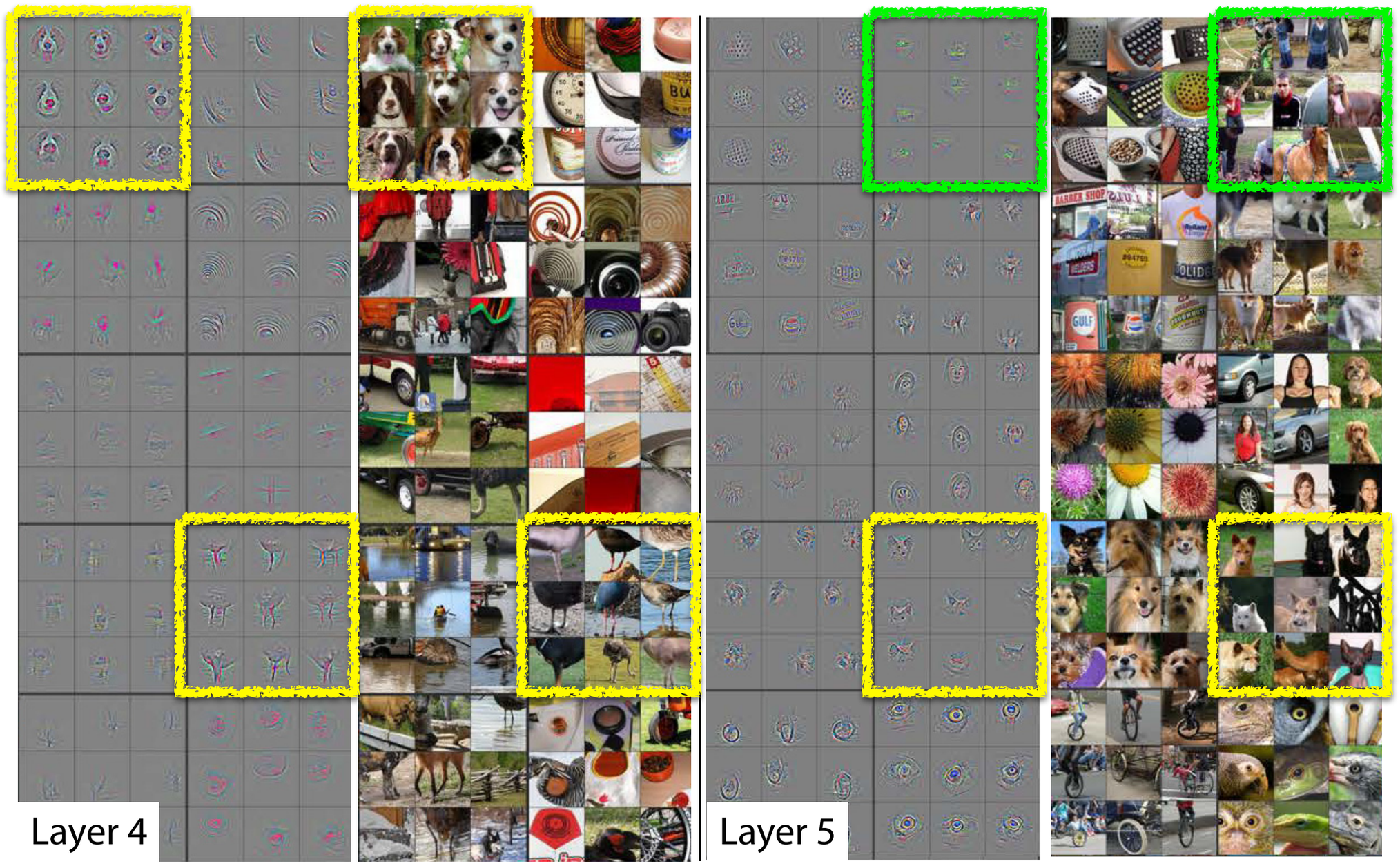
Feature visualization



Object parts
(dog face & bird legs)

Entire object with pose variation
(dogs)

Feature visualization

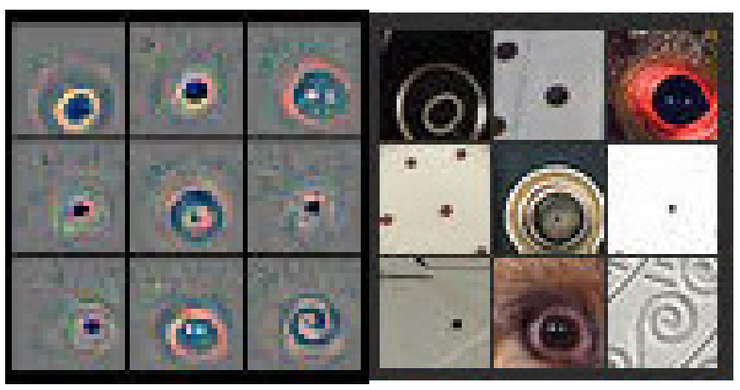


Object parts
(dog face & bird legs)

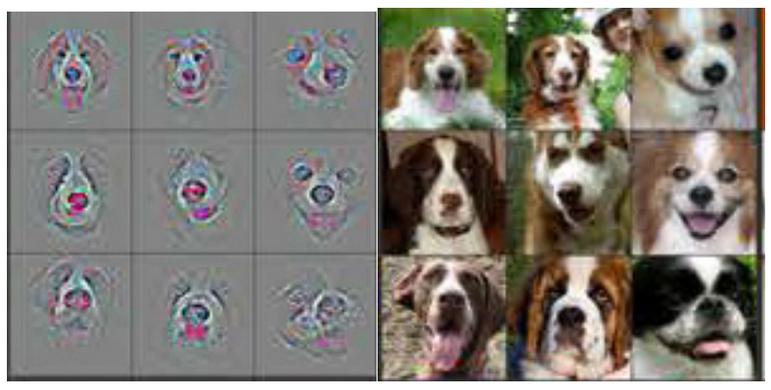
Entire object with pose variation
(dogs)

Notes

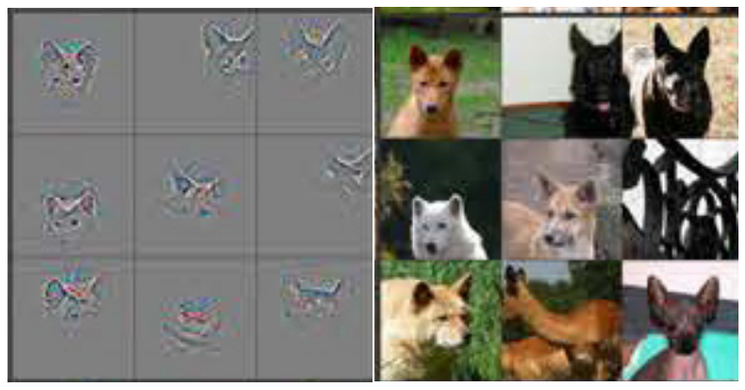
- Hierarchical representation of features
- Strong grouping within each feature map
- Larger invariance in higher layers (Layer 5)
- Selection of discriminative parts of images



Layer 2

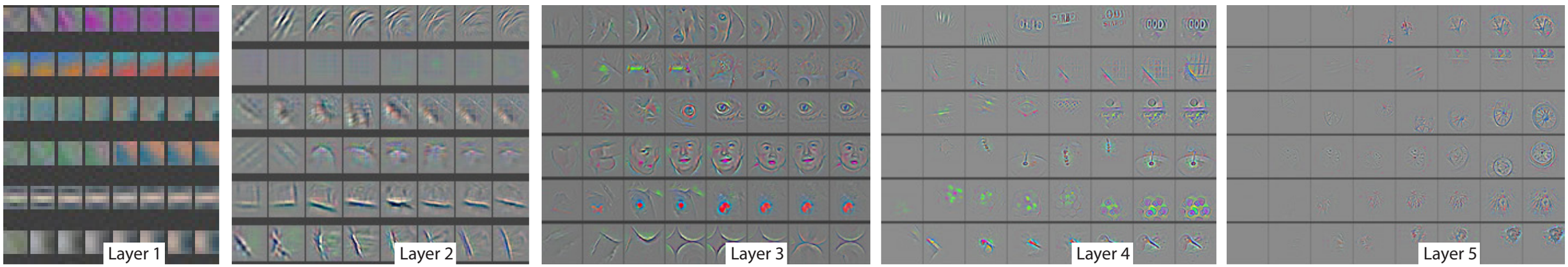


Layer 4



Layer 5

Feature evolution during training

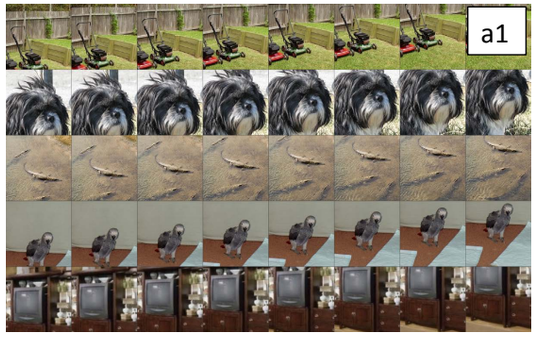


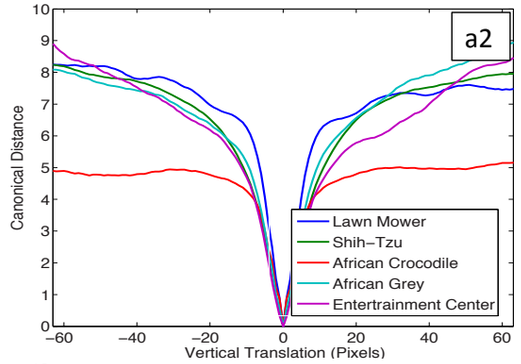
- Lower layers converge faster
- Higher layers start to converge later
- Sudden jump: different images result strong activation

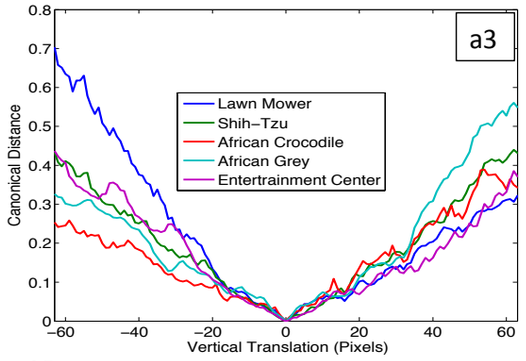
Feature invariance

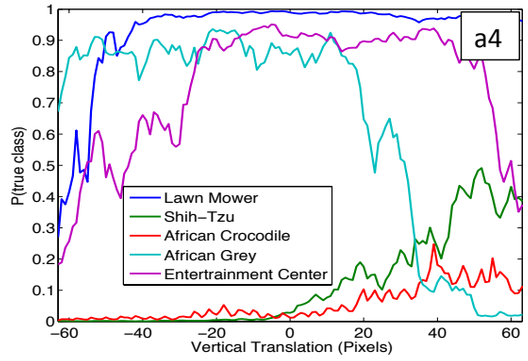
Euclidean distance between feature of transformed and original

Translation

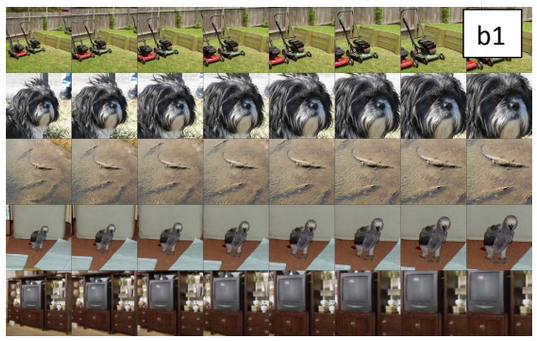


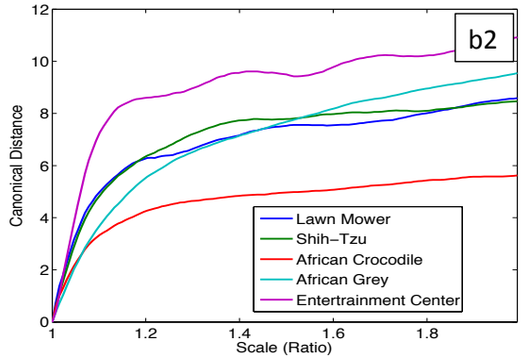


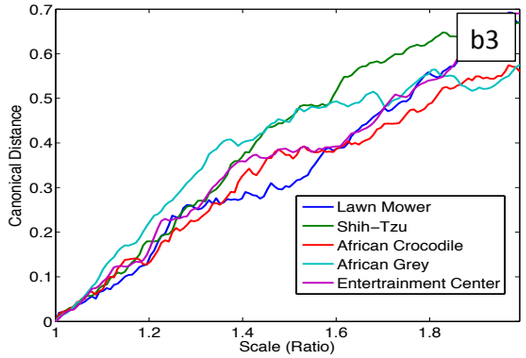


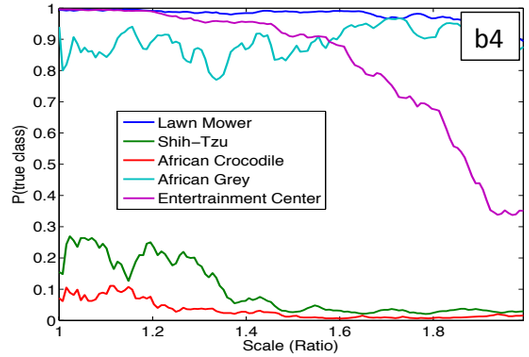


Scale

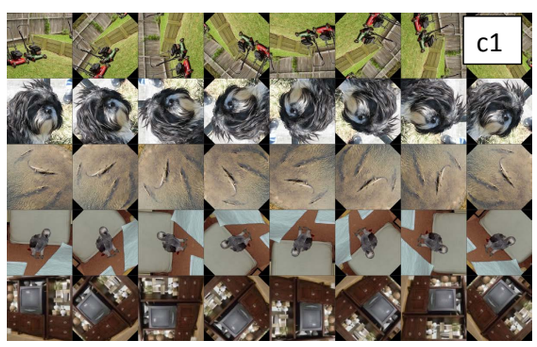


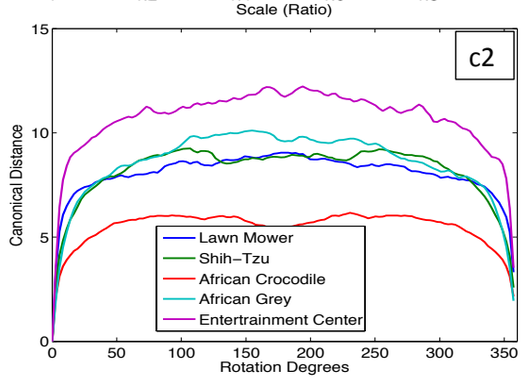


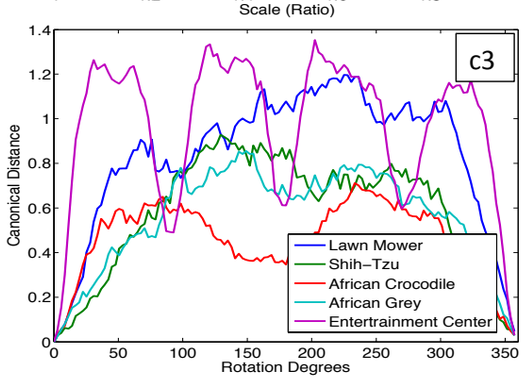


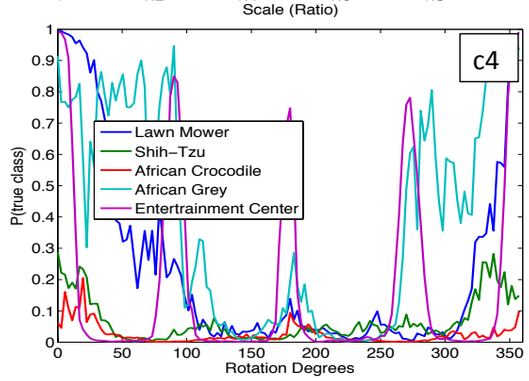


Rotation









drastic change

Layer 1

quasi-linear

Layer 7

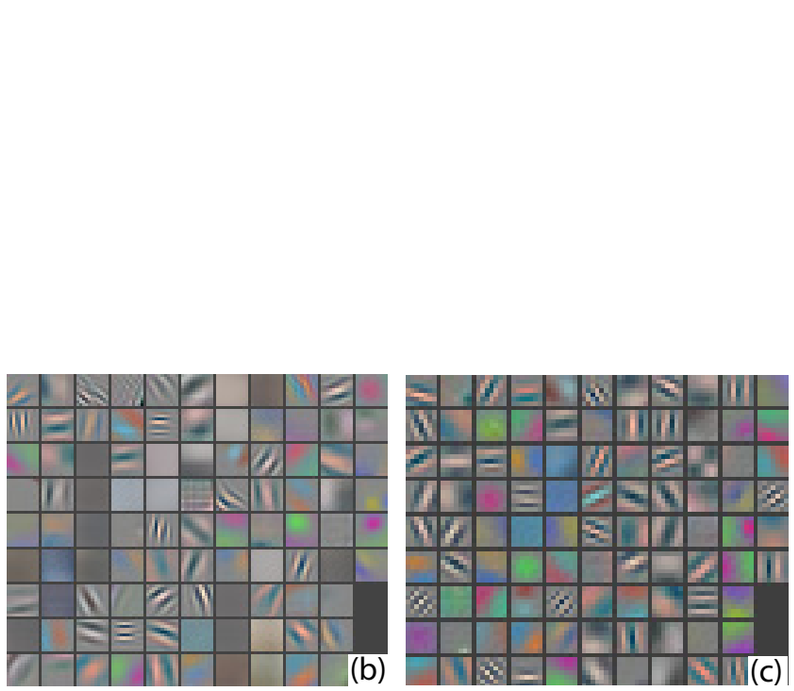
invariant

Output layer

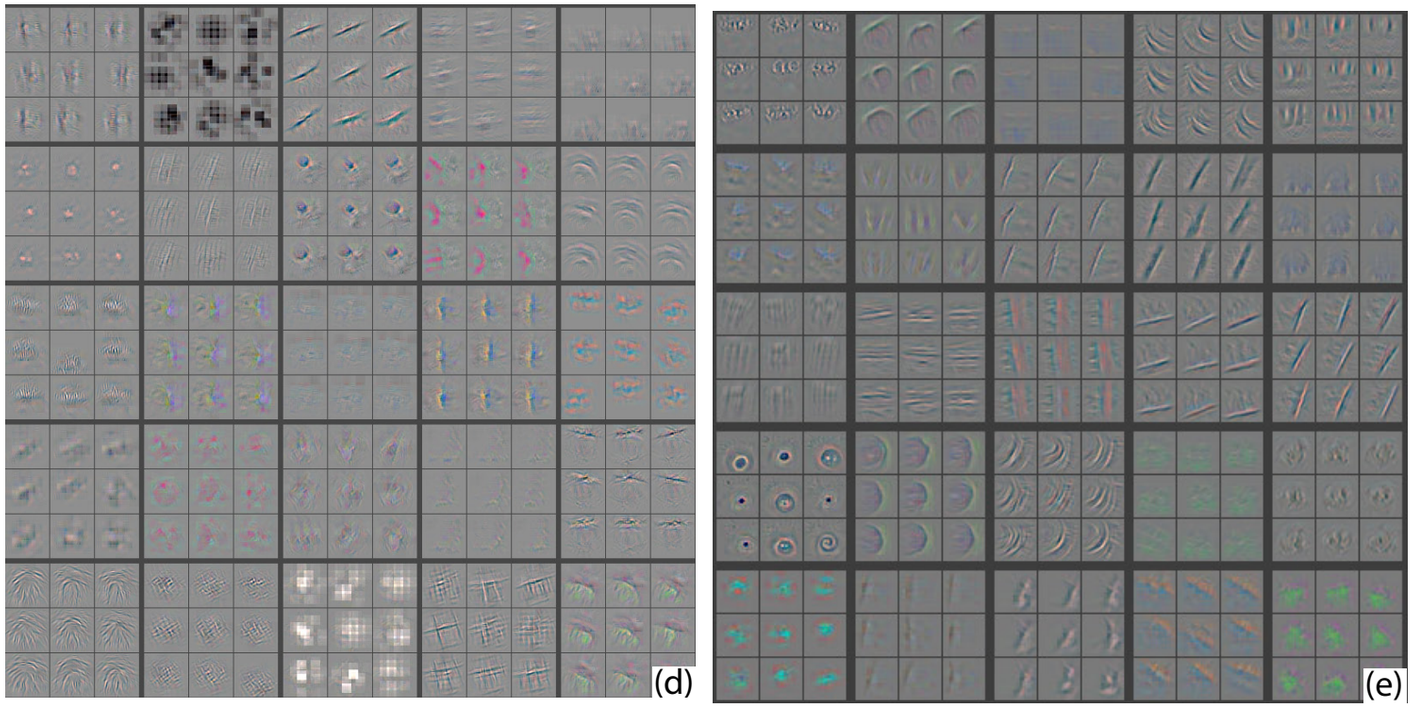
Feature layers

Output layer

Architecture selection



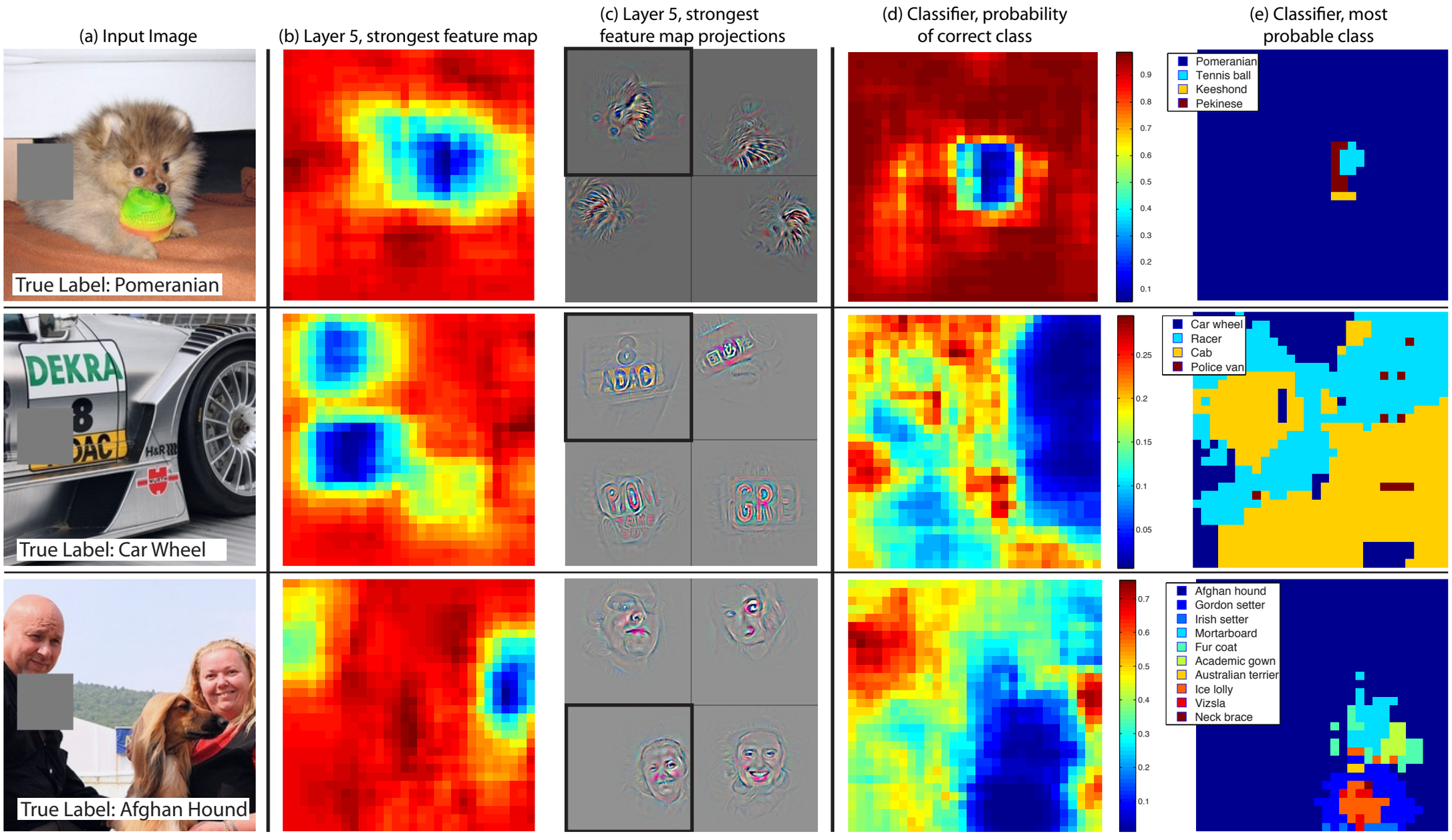
Layer 1



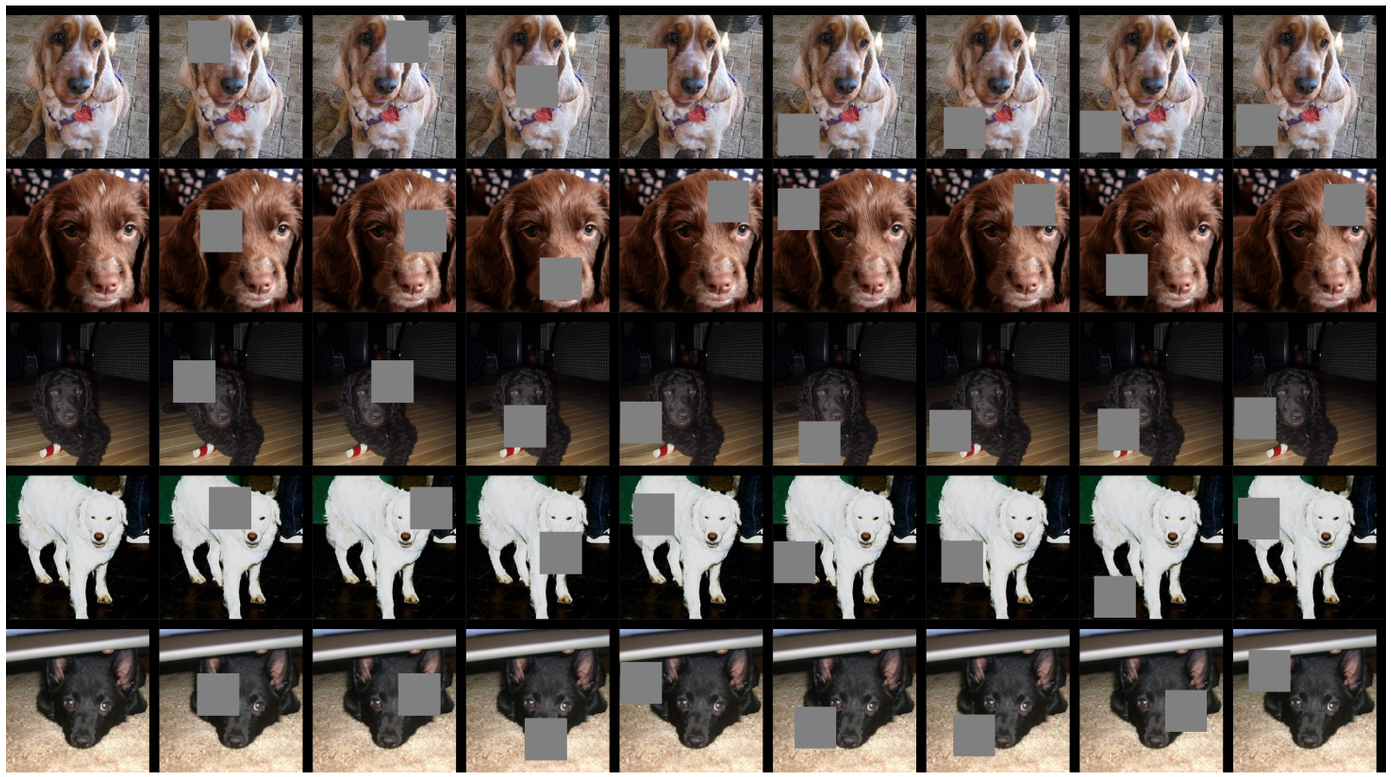
Layer 2

- 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

Occlusion sensitivity



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)

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

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