Visualizing and Understanding Convolution Networks

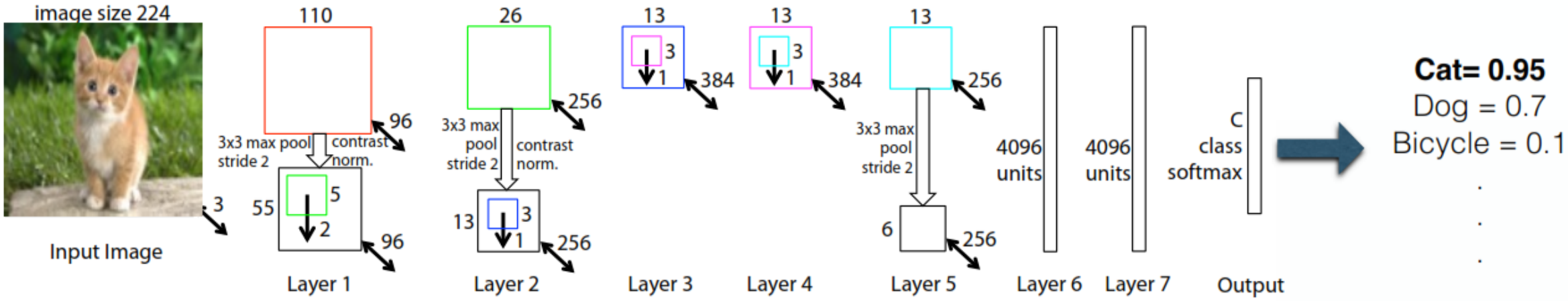
Authors: Mathew Zeiler and Rob Fergus
New York University
Select slides from Hamid Izadinia

Presentation by Jason Driver

10/13/2016

Problem and Contributions

- Greatly improved performance using convolution neural network architecture over the previous state of the art
- No clear understanding why CNN work
- Technique for visualizing individual layers with a Deconvolution Network
- Troubleshoot the output of each layer
- Introduces an image occlusion experiment for spatial understanding
- Optimal architecture



Hierarchical Convolution

Nonlinear operations (max pooling, ReLu)

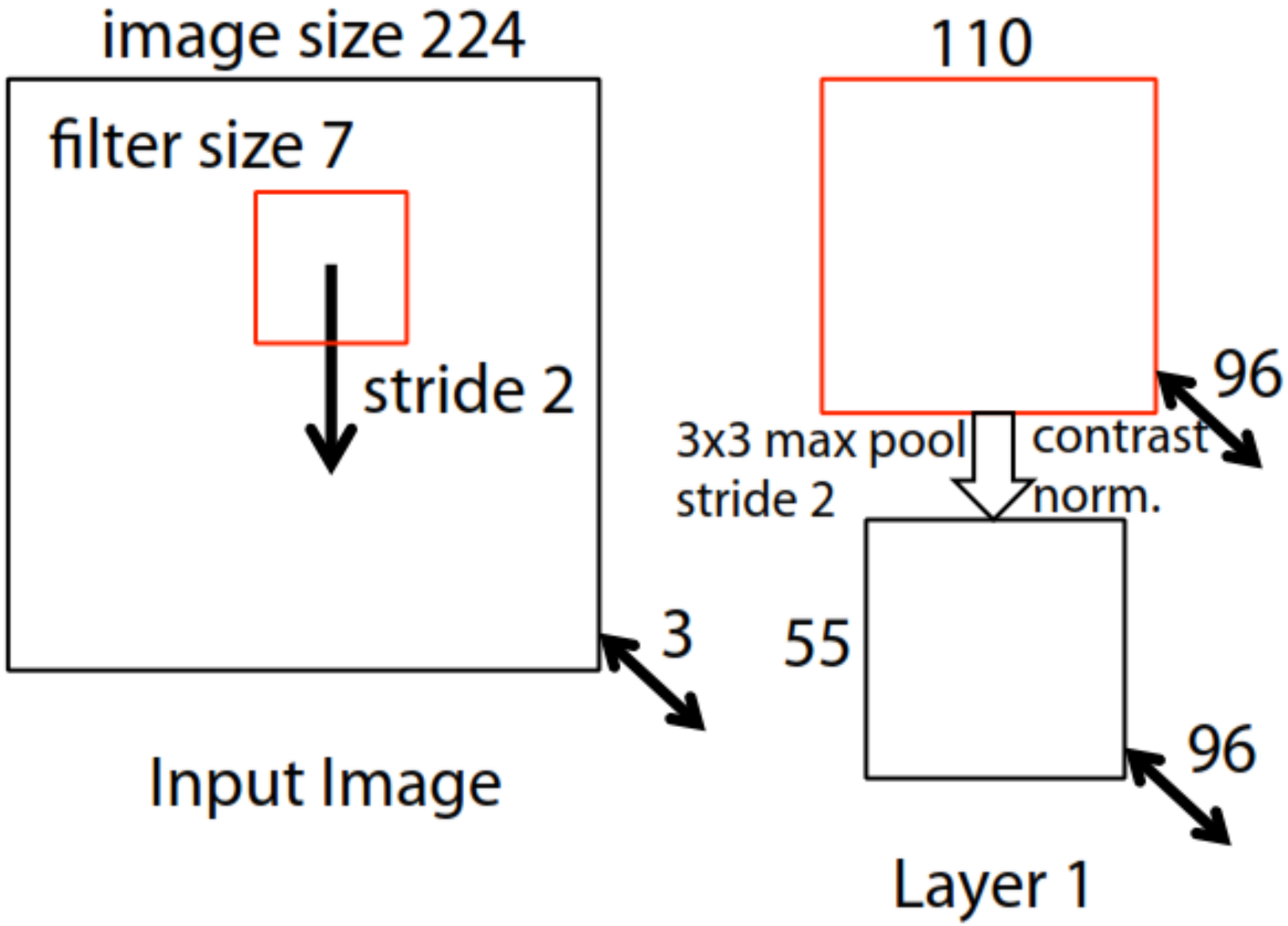


image size 224
filter size 7
stride 2
3
Input Image
110
96
3x3 max pool
stride 2
contrast
norm.
55
96
Layer 1

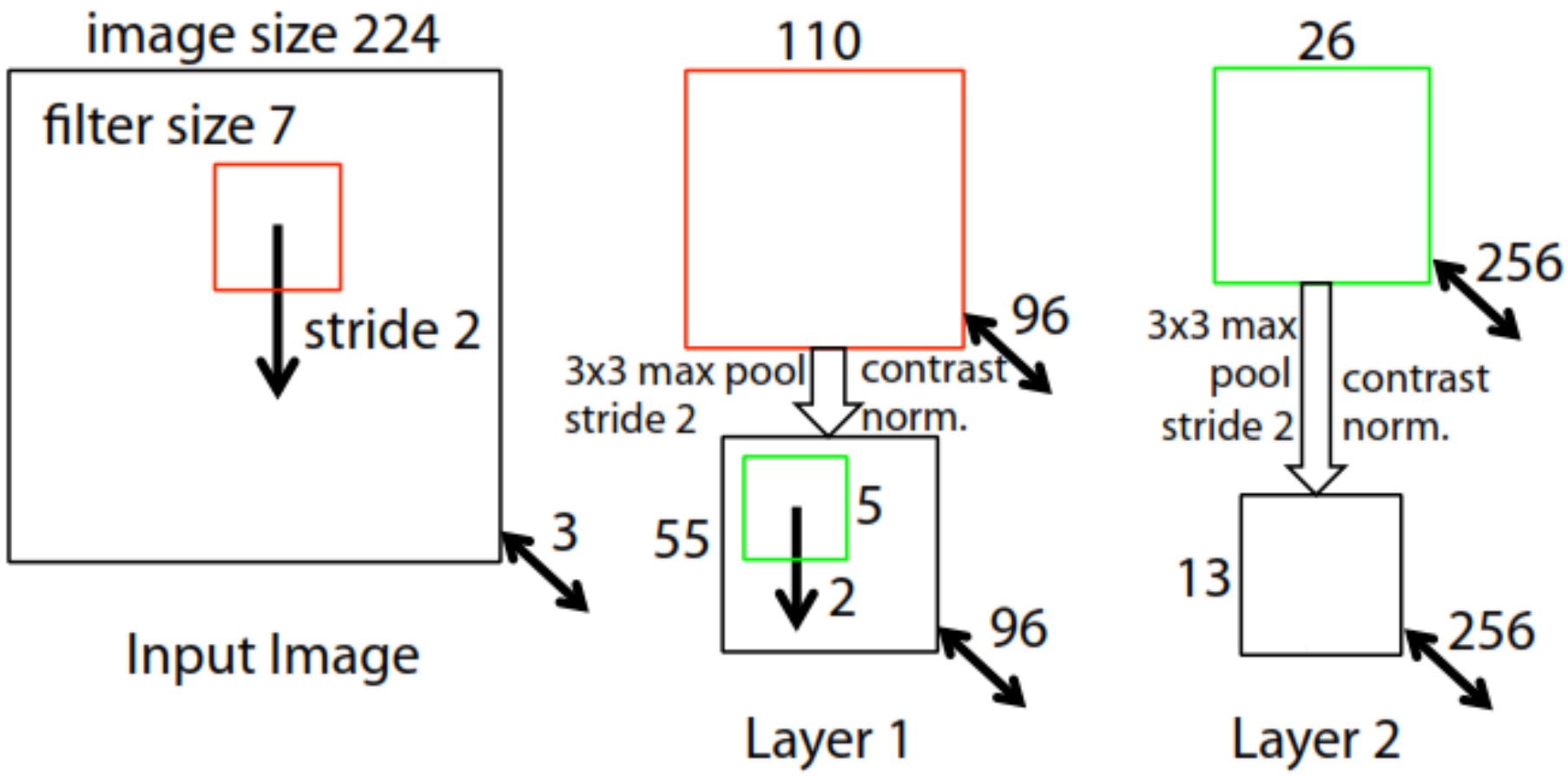
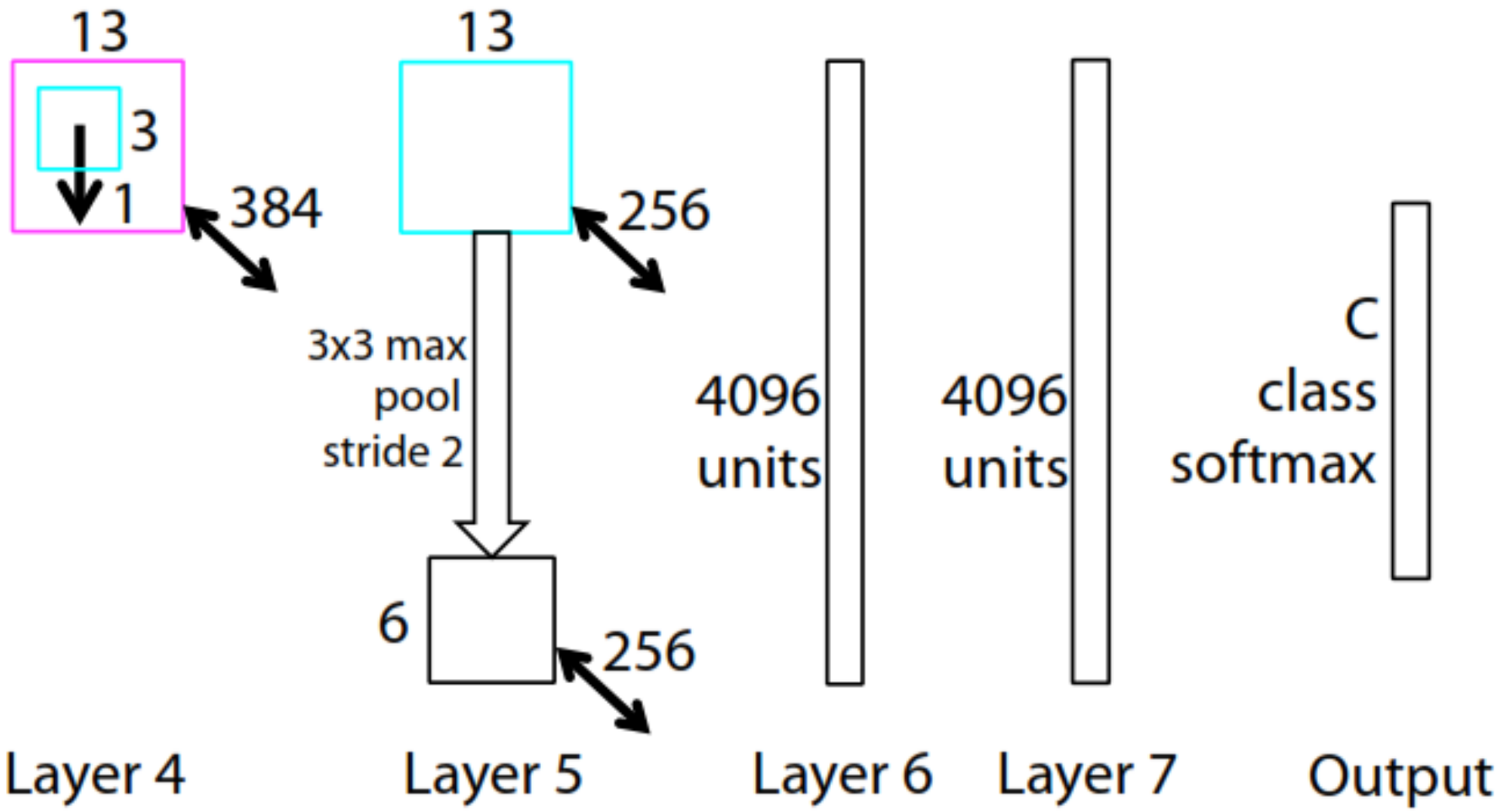
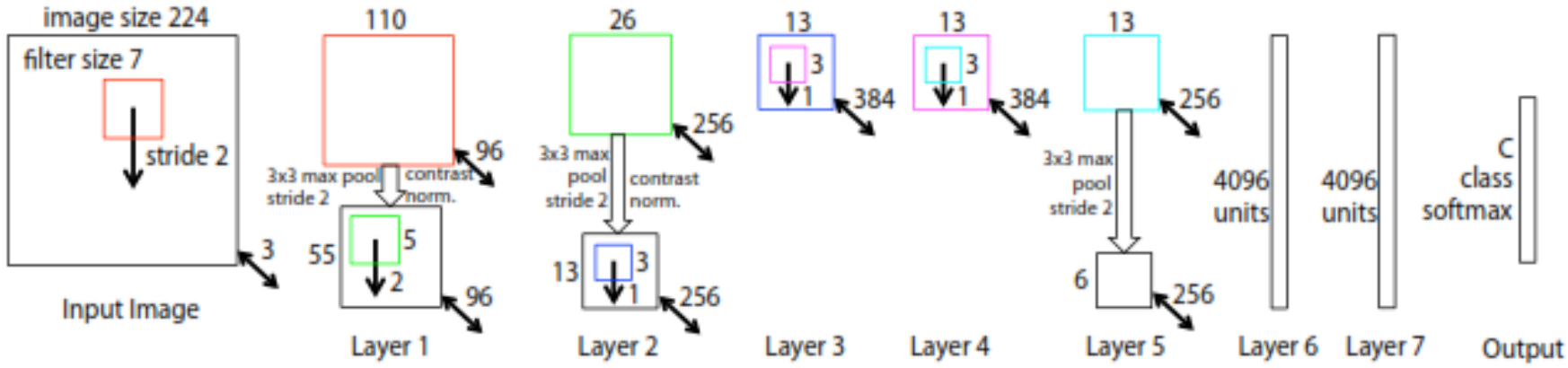


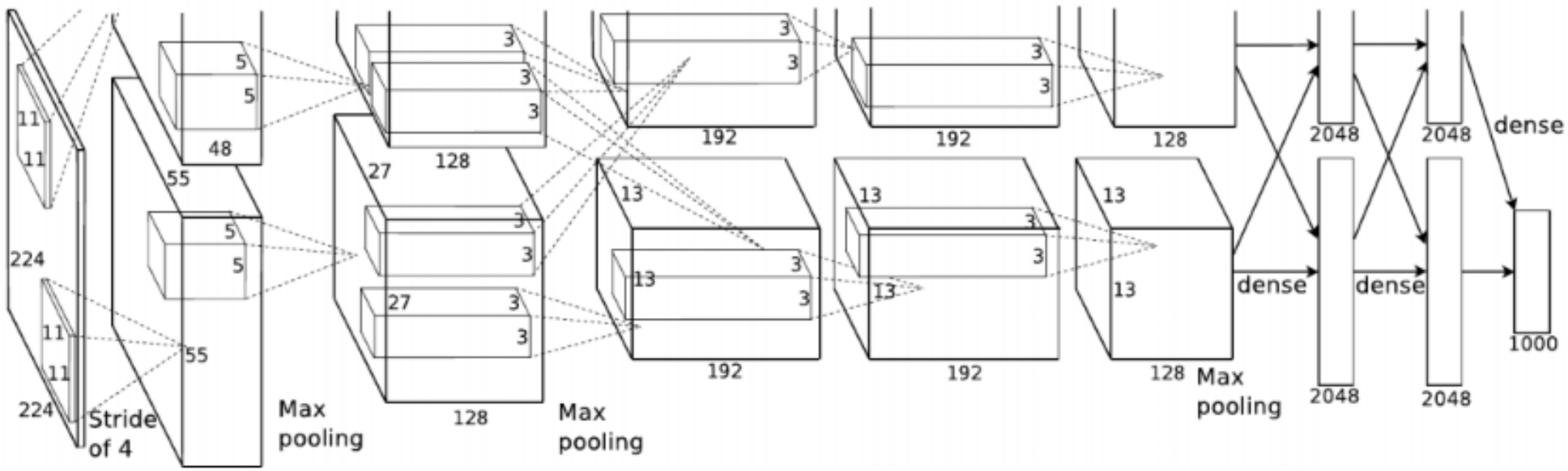
image size 224
filter size 7
stride 2
3
Input Image
110
96
3x3 max pool
stride 2
contrast
norm.
55
5
2
96
Layer 1
26
256
3x3 max
pool
stride 2
contrast
norm.
13
256
Layer 2


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

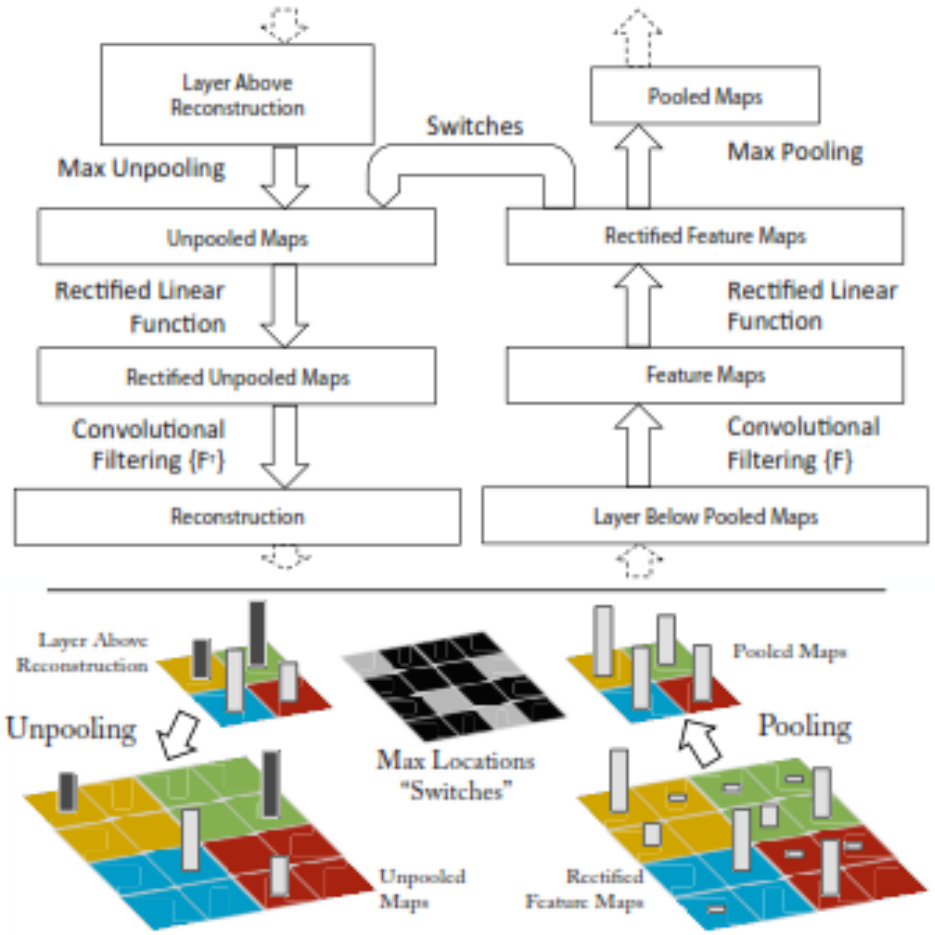
Convolutional Neural Network



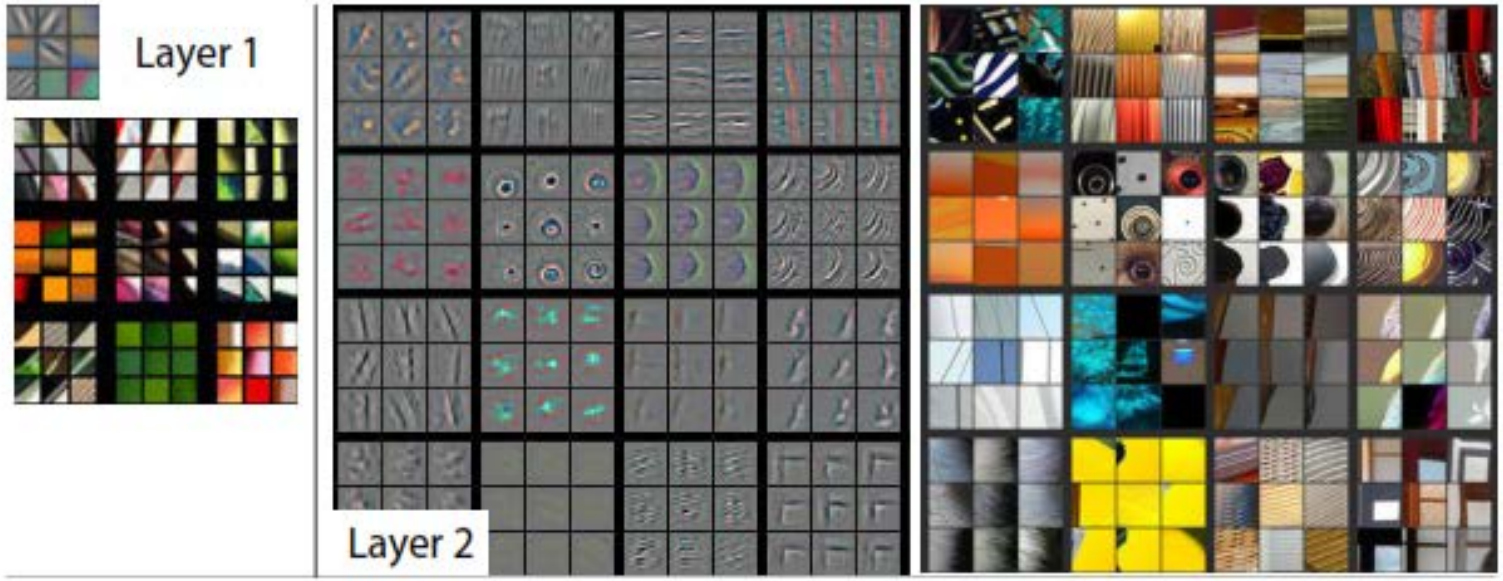
AlexNet


224
224
11
11
11
11
Stride
of 4
55
55
48
5
5
5
5
Max
pooling
27
27
128
128
3
3
3
3
3
3
3
3
Max
pooling
13
13
13
13
192
192
3
3
3
3
3
3
192
192
13
13
13
13
128
128
Max
pooling
dense
dense
dense
2048
2048
2048
2048
1000

Deconvnet & Convnet

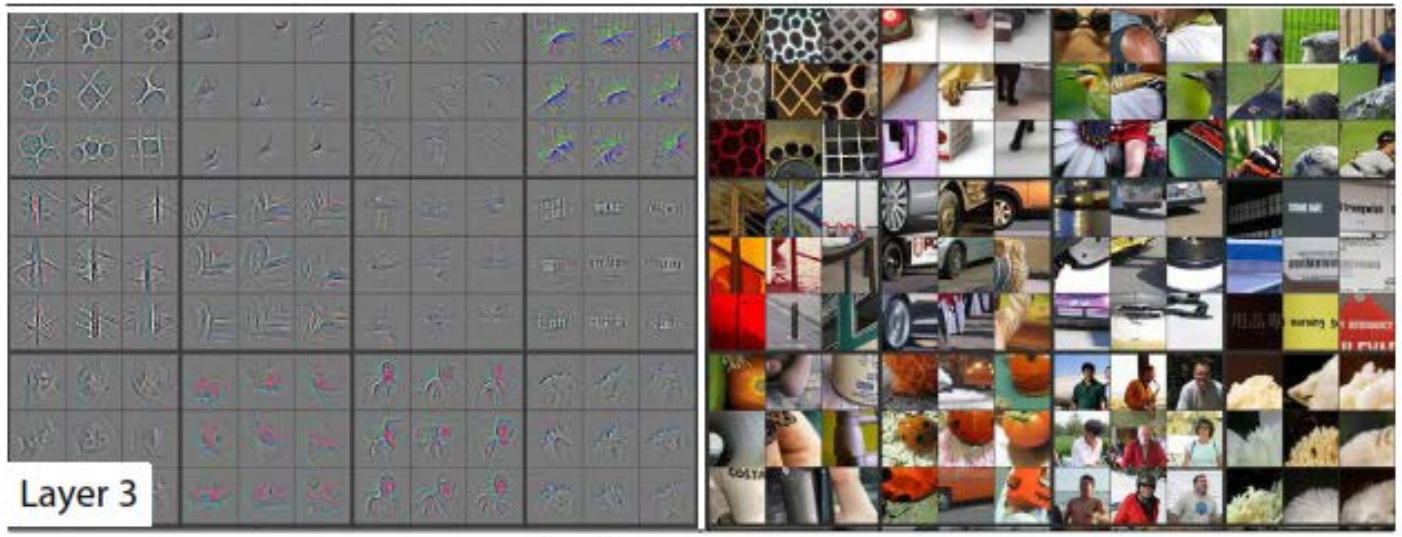


Feature Visualization



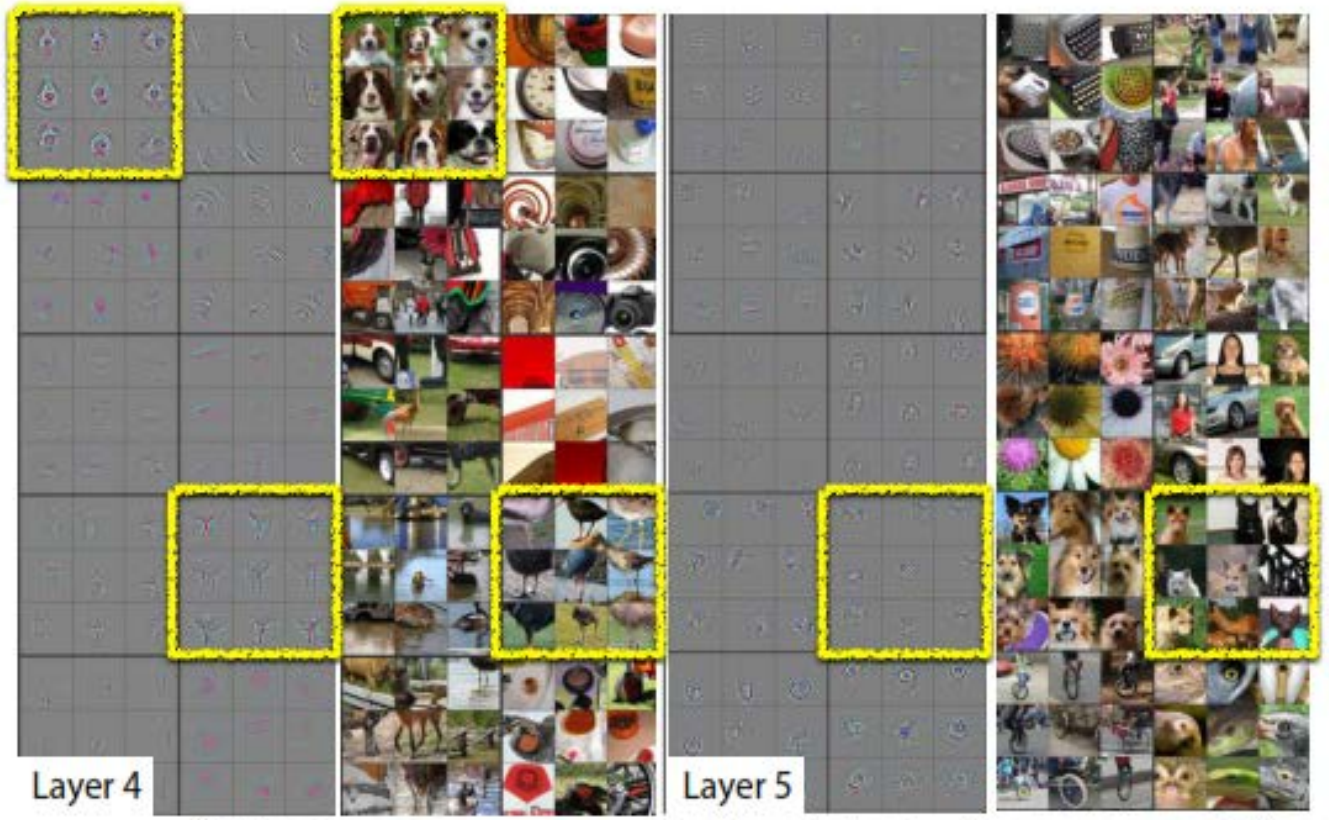
corners & edge/color conjunctions

Feature Visualization



similar textures

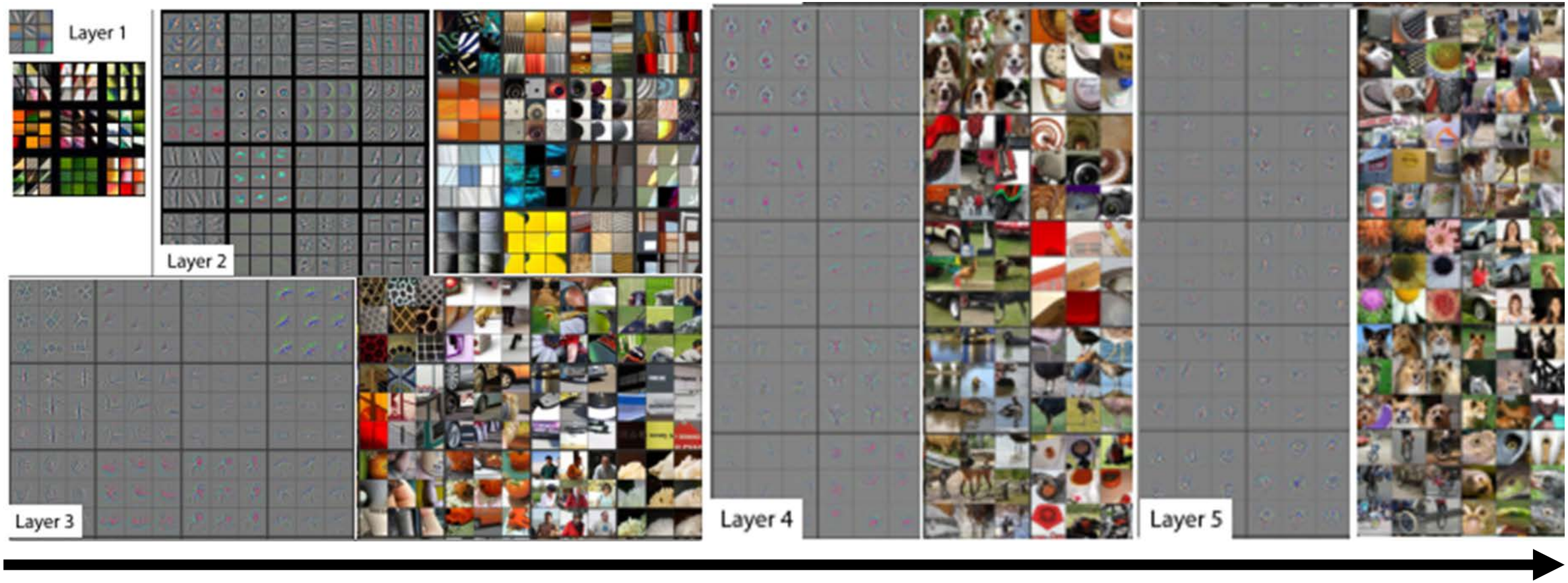
Feature Visualization



Object parts
(dog face & bird legs)

Entire object with pose variation
(dogs)

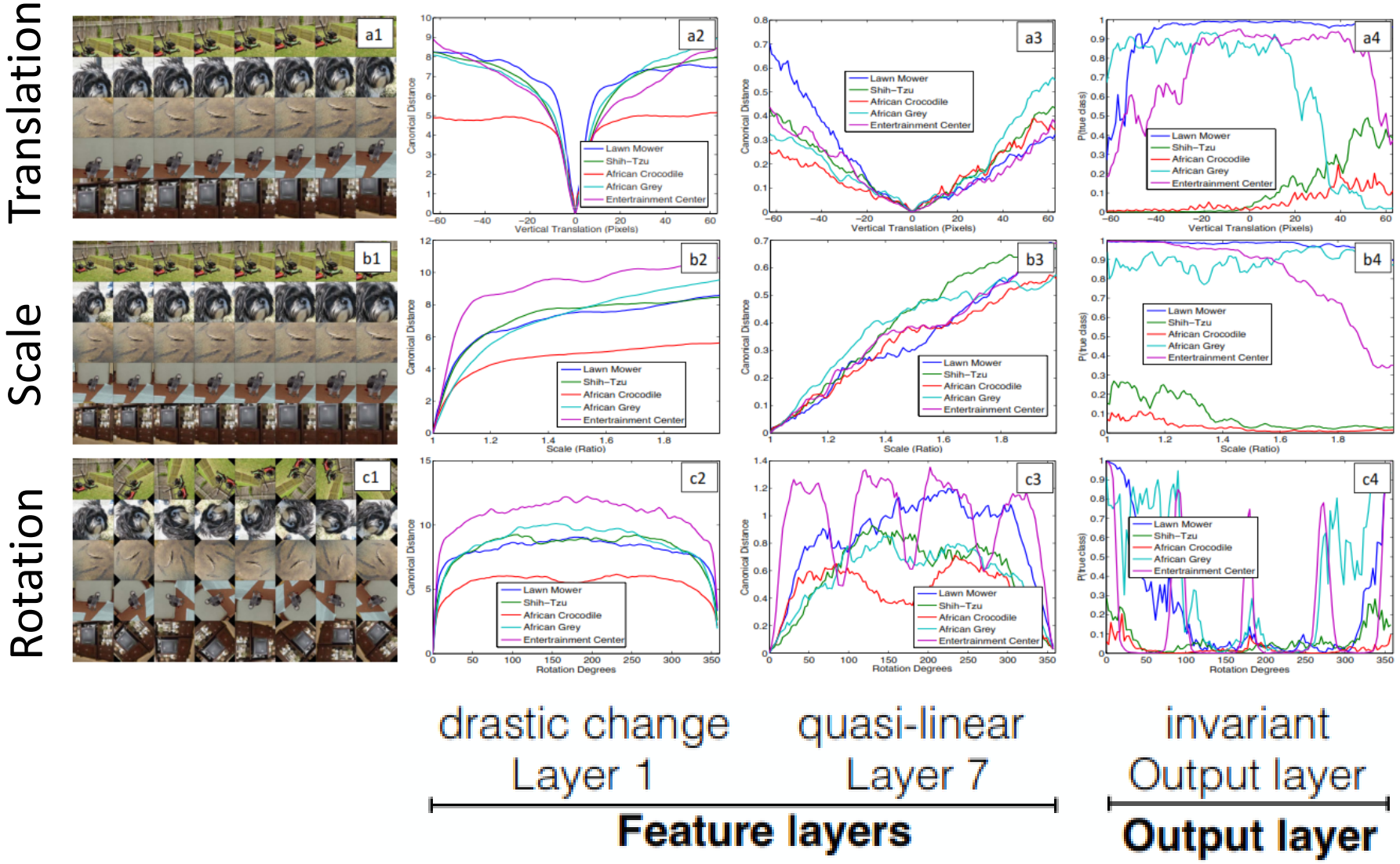
Feature Visualization



Lower Order Features

Higher Order Features

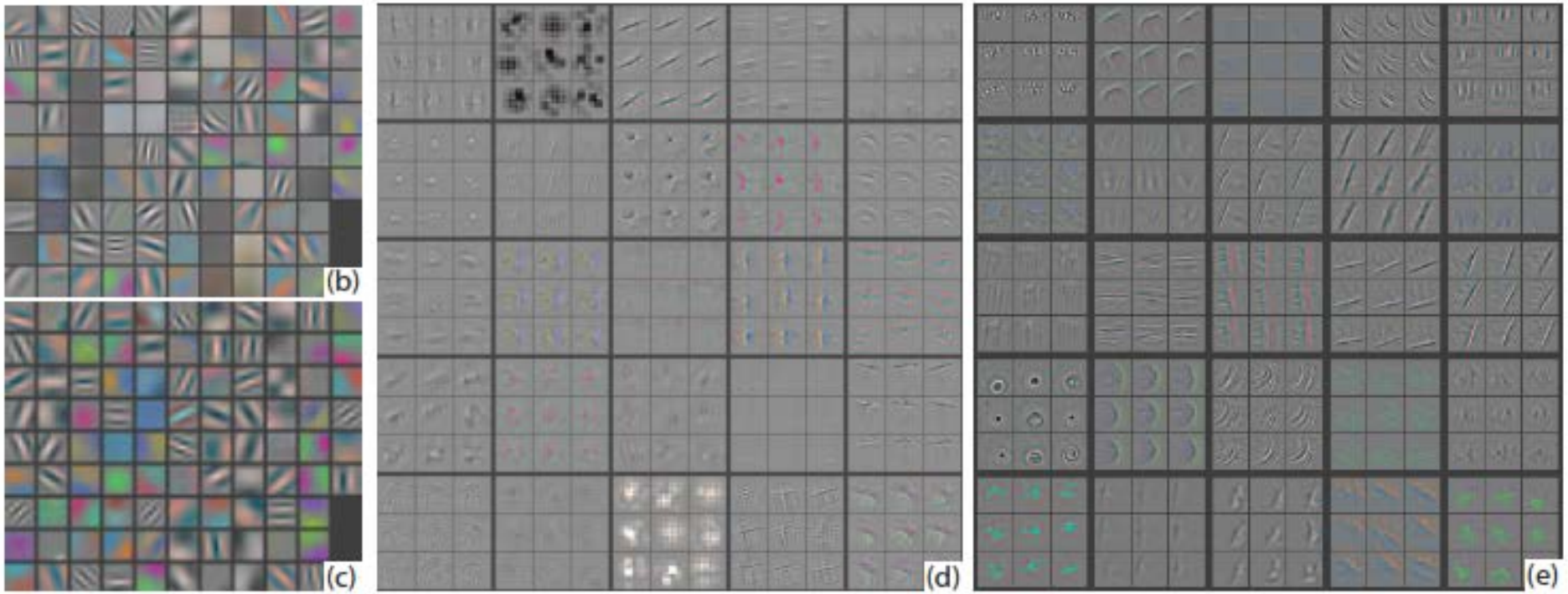
Feature Invariance in Layers 1, 7, 8



$$\epsilon_i^l = x_i^l - \tilde{x}_i^l$$

Euclidean distance between feature of transformed and original

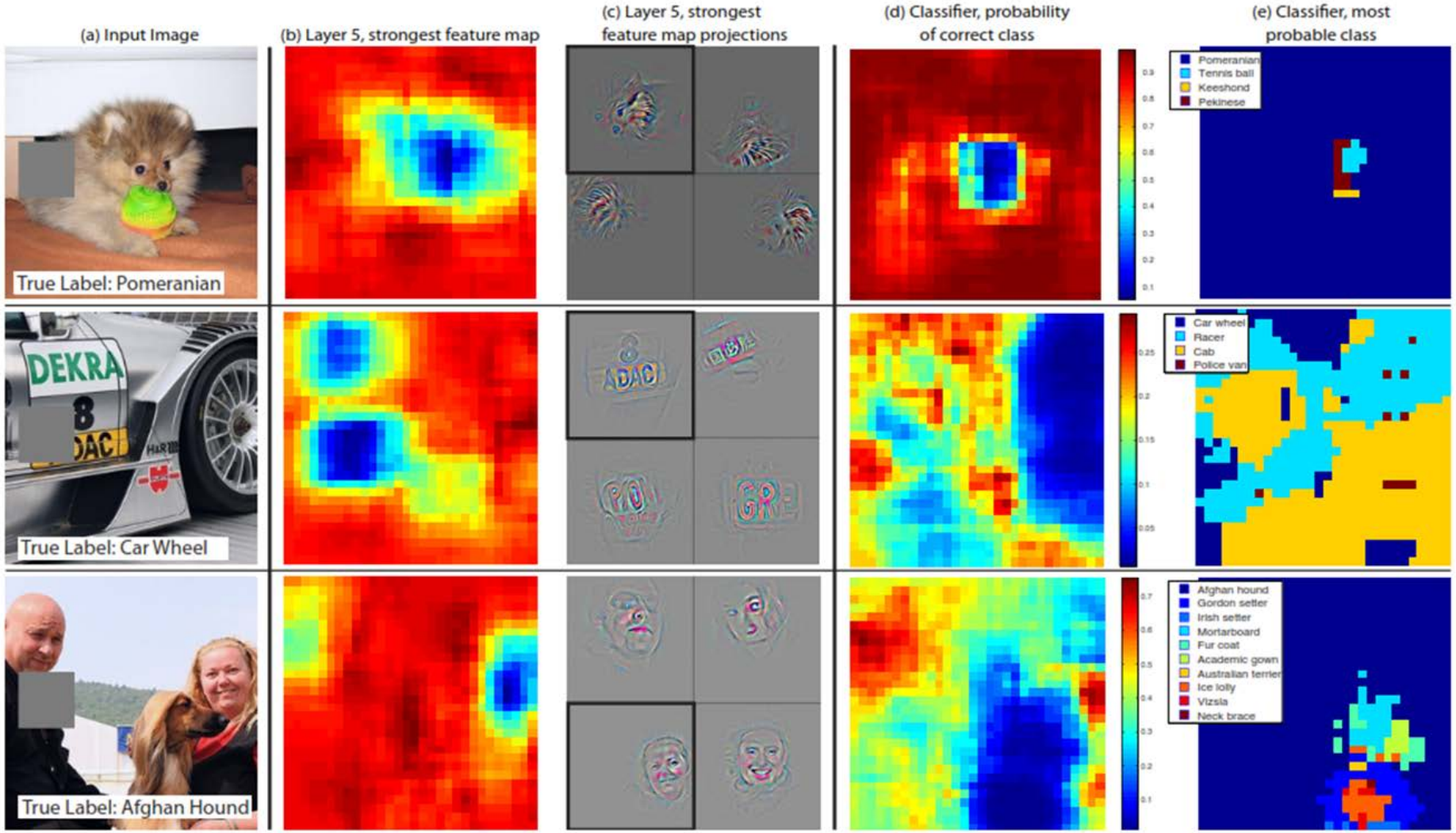
Architecture Selection



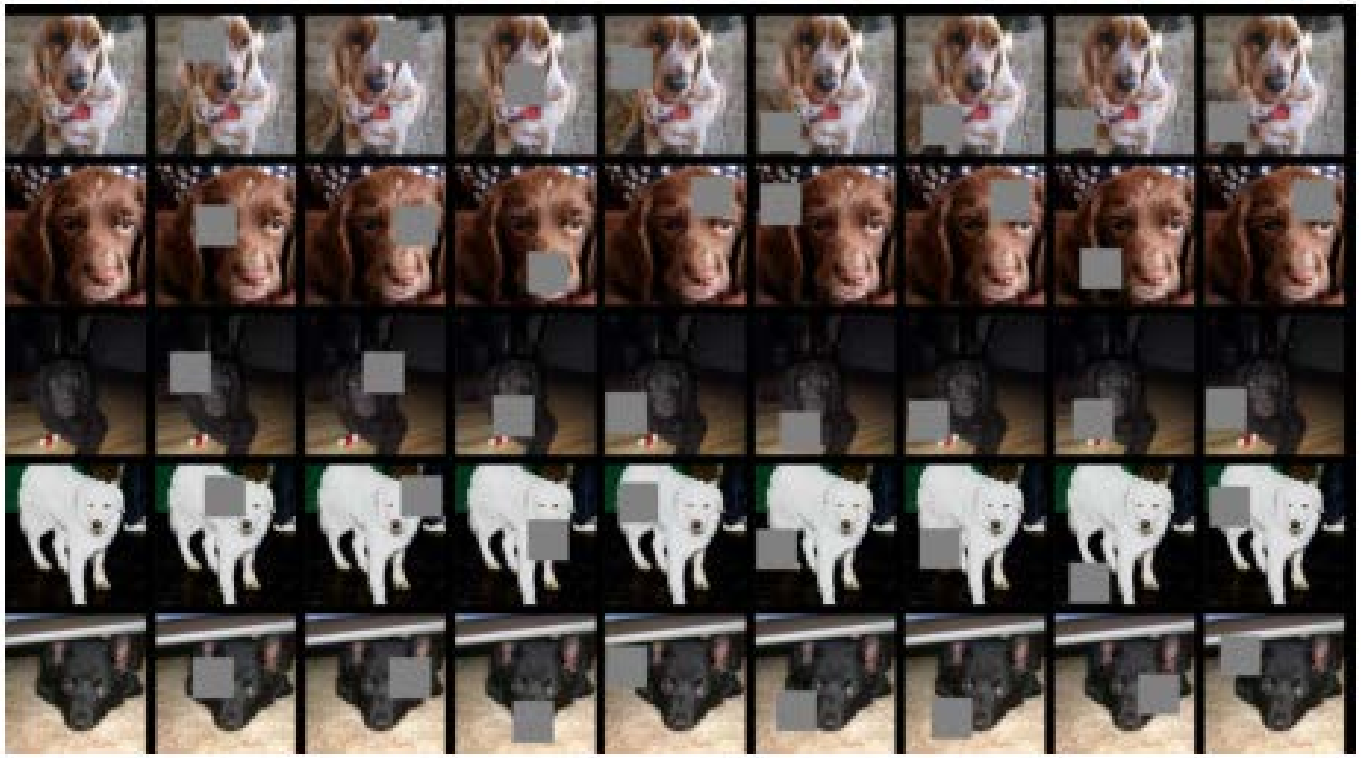
Layer 1 Layer 2

- Smaller filters (7x7 vs. 11x11) and smaller stride (2 vs. 4)
- 1st Layer: Increased coverage of mid-frequencies (Original is b)
- 2nd Layer: no aliasing, no “dead” features (Original is d)

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

- 5th Layer preserves correspondence
- 7th Layer discriminates different classes of dog

Results

Error %	Val Top-1	Val Top-5	Test Top-5
(Gunji et al., 2012)	-	-	26.2
(Krizhevsky et al., 2012), 1 convnet	40.7	18.2	−−
(Krizhevsky et al., 2012), 5 convnets	38.1	16.4	16.4
(Krizhevsky et al., 2012)*, 1 convnets	39.0	16.6	−−
(Krizhevsky et al., 2012)*, 7 convnets	36.7	15.4	15.3
Our replication of (Krizhevsky et al., 2012), 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**