Statistical Methods in AI (CSE 471)
Lecture16: CNN

Vineet Gandhi
Centre for Visual Information Technology (CVIT)

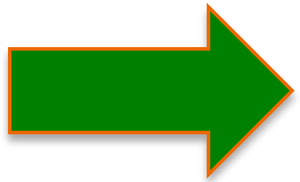
Difficult problem

194	210	201	212	199	213	215	195	178	158	182	209
180	189	190	221	209	205	191	167	147	115	129	163
114	126	140	188	176	165	152	140	170	106	78	88
87	103	115	154	143	142	149	153	173	101	57	57
102	112	106	131	122	138	152	147	128	84	58	66
94	95	79	104	105	134	129	113	107	87	69	67
68	71	69	98	89	92	98	95	89	88	76	67
41	56	68	99	83	45	60	83	58	76	75	65
20	43	69	75	56	41	51	73	55	70	63	44
50	50	57	69	75	75	73	74	53	68	59	37
72	59	53	66	84	92	84	74	57	72	63	42
87	61	58	65	75	78	76	73	59	75	69	50

Cow vs Zebra

Cow vs Zebra

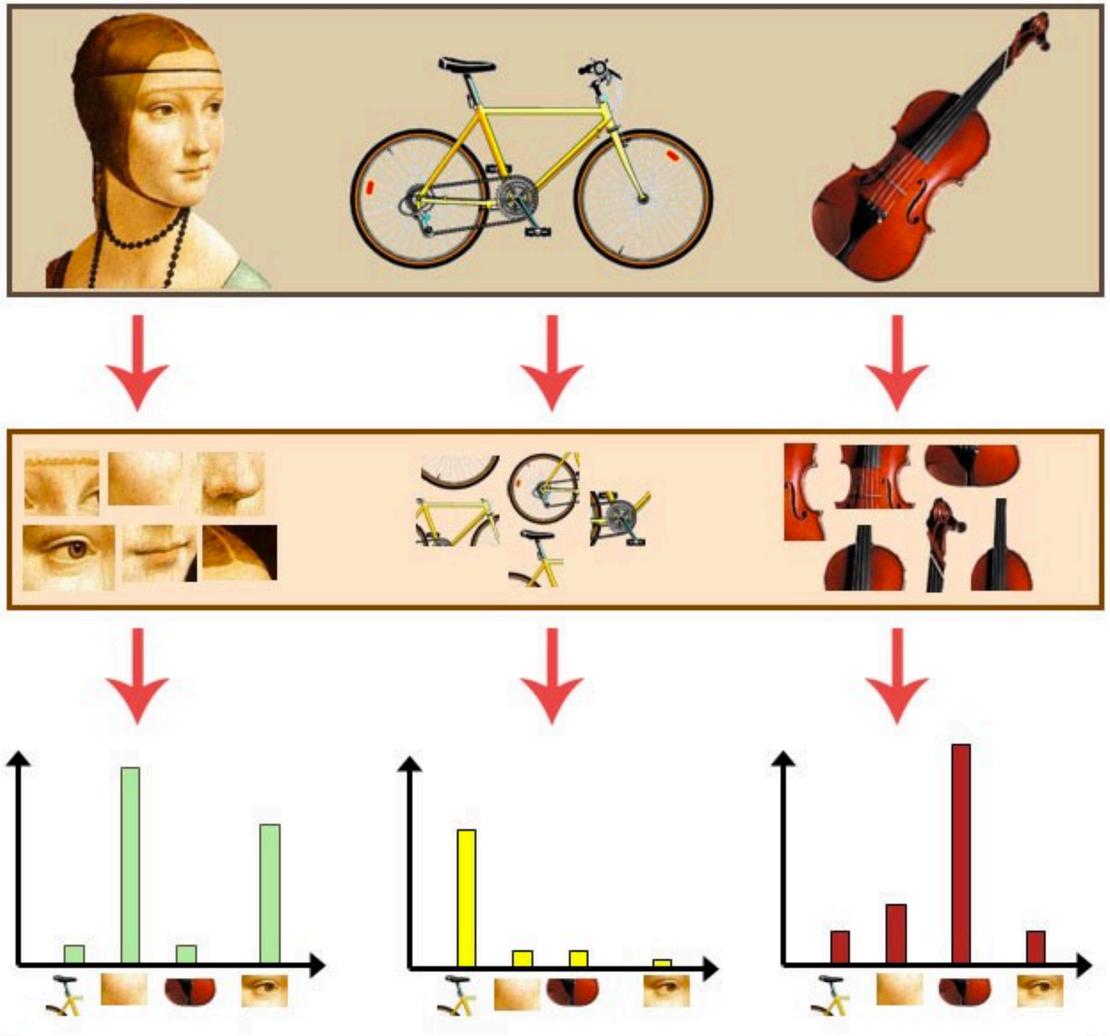
Raw image?



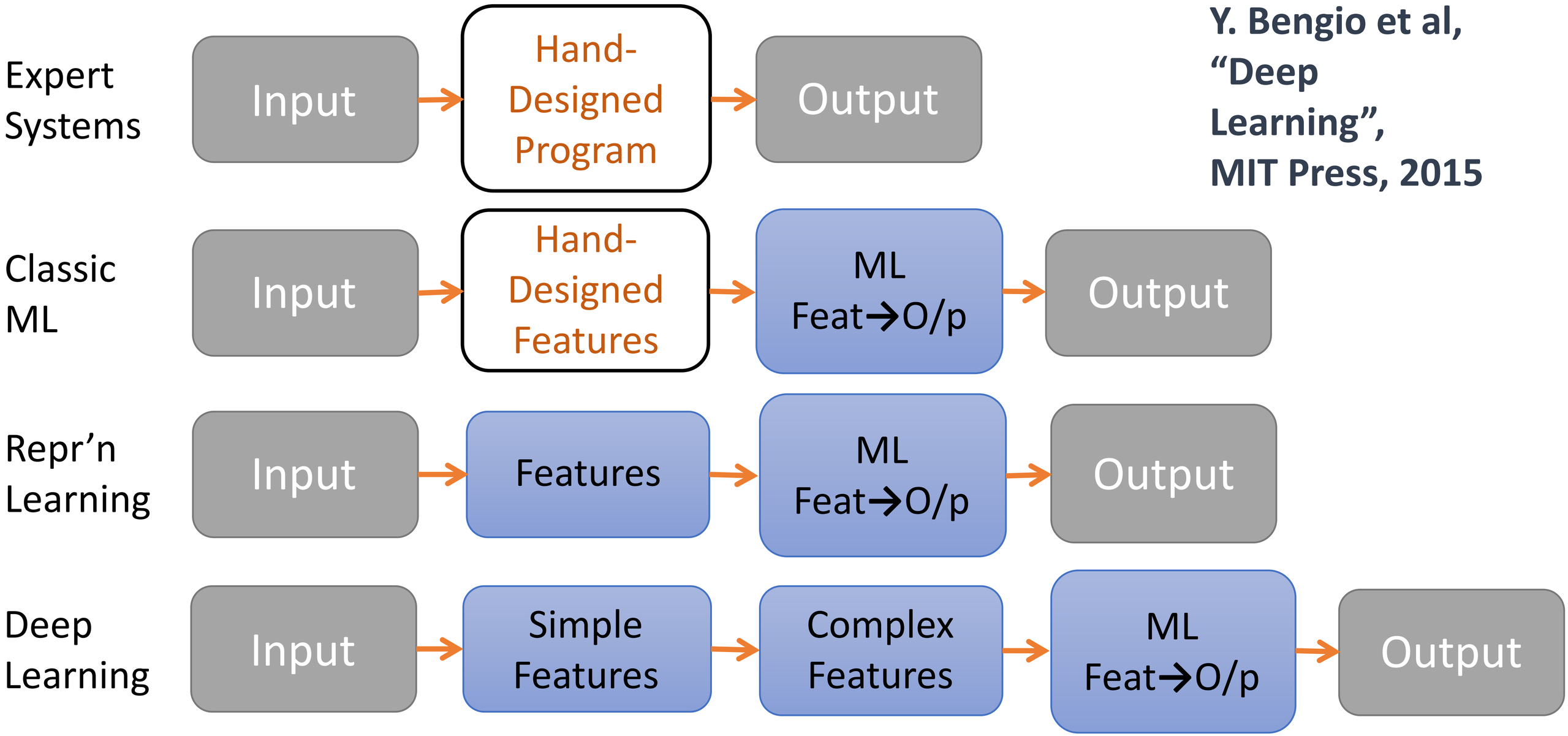
Other features

MIN RED
MAX RED
MEAN RED
MIN GREEN
MAX GREEN
MEAN GREEN
MIN BLUE
MAX BLUE
MEAN BLUE

Higher level representations: bag of words

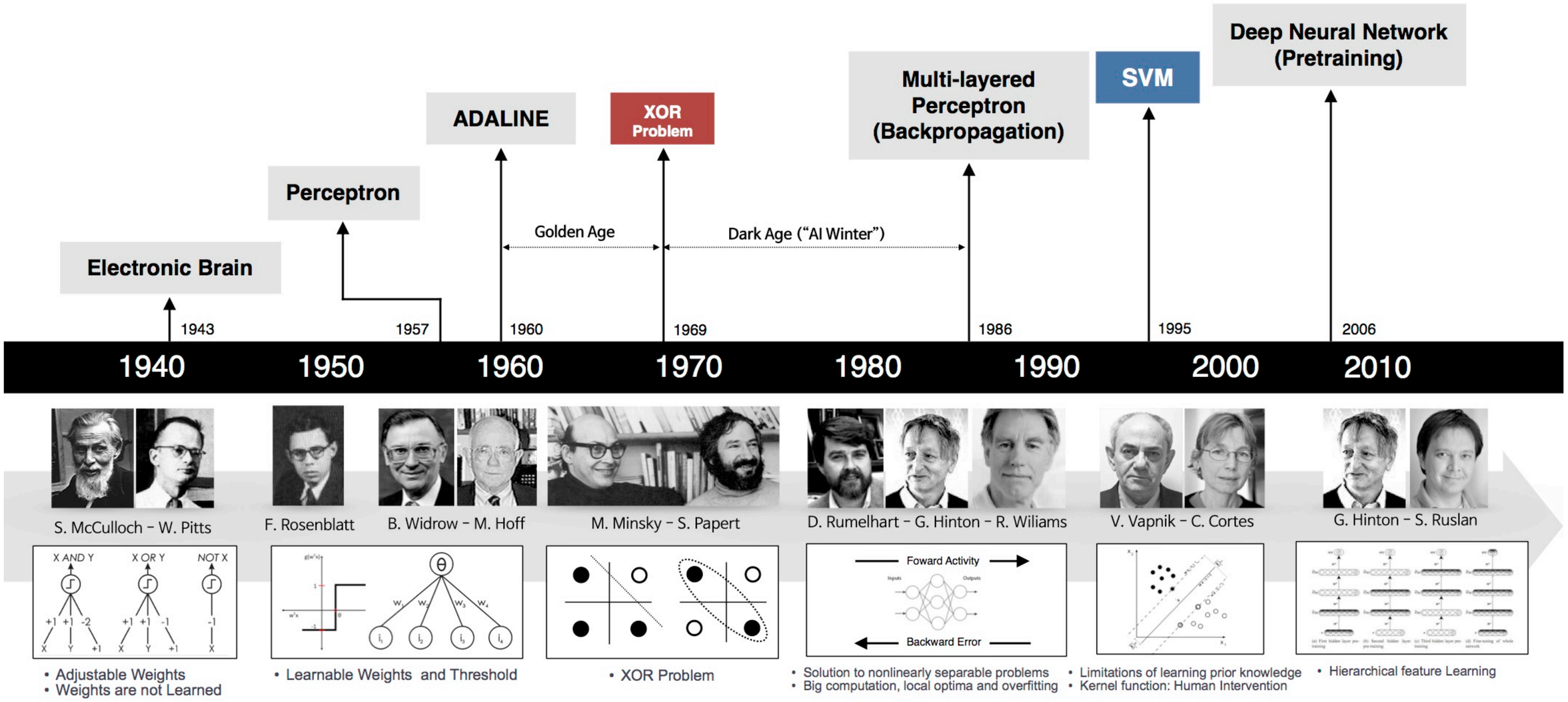


CNN



Y. Bengio et al, "Deep Learning", MIT Press, 2015

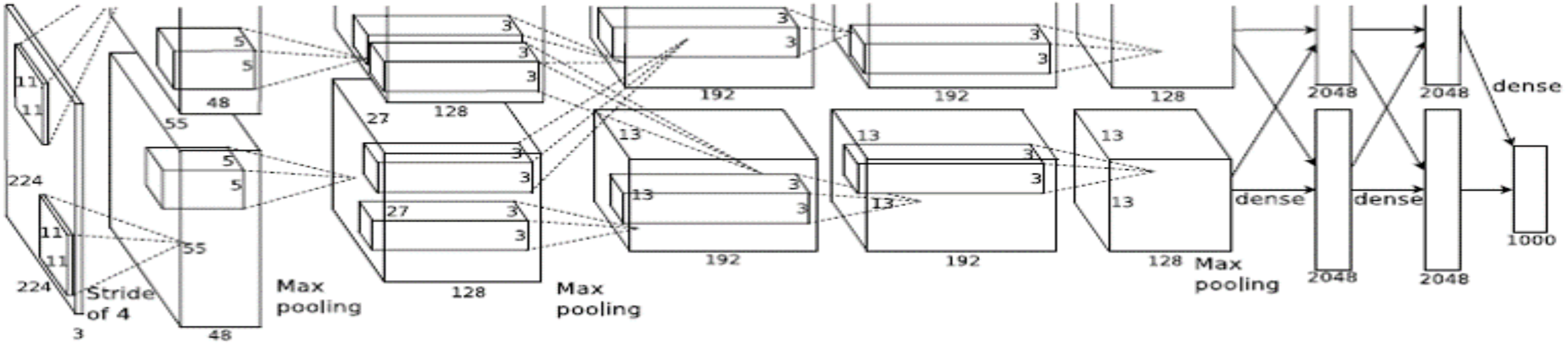
History


Deep Neural Network (Pretraining)
SVM
Multi-layered Perceptron (Backpropagation)
XOR Problem
ADALINE
Perceptron
Electronic Brain
Golden Age
Dark Age ("AI Winter")
1943
1957
1960
1969
1986
1995
2006
1940
1950
1960
1970
1980
1990
2000
2010
S. McCulloch - W. Pitts
F. Rosenblatt
B. Widrow - M. Hoff
M. Minsky - S. Papert
D. Rumelhart - G. Hinton - R. Wiliams
V. Vapnik - C. Cortes
G. Hinton - S. Ruslan
X AND Y
X OR Y
NOT X
Foward Activity
Backward Error
• Adjustable Weights
• Weights are not Learned
• Learnable Weights and Threshold
• XOR Problem
• Solution to nonlinearly separable problems
• Big computation, local optima and overfitting
• Limitations of learning prior knowledge
• Kernel function: Human Intervention
• Hierarchical feature Learning

ImageNet

IMAGENET

CNN



ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

ImageNet Classification Task:

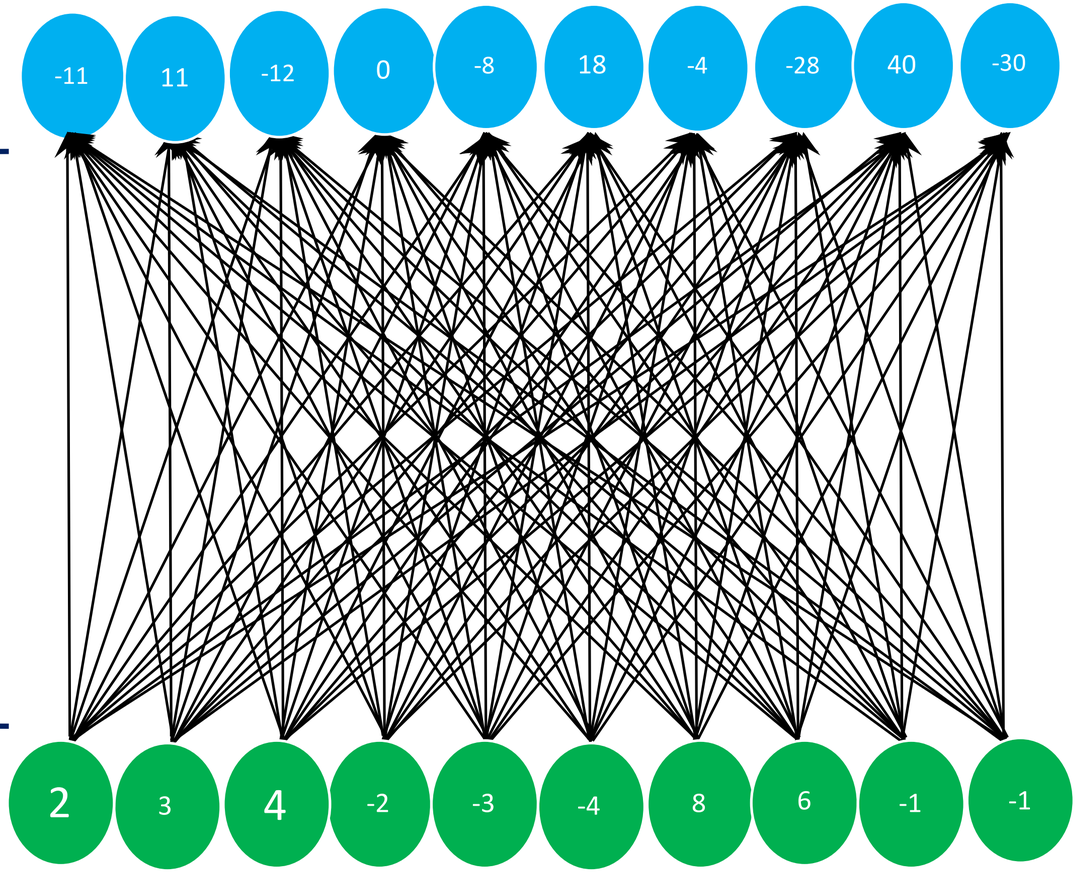
Previous Best : ~25% (CVPR-2011)
AlexNet : ~15 % (NIPS-2012)

Dense connections

$$Y = W^T X$$

$$W = \begin{bmatrix} -1 & 1 & -2 & 2 & -1 & 1 & 2 & -2 & 3 & -3 \\ -2 & 2 & -1 & 1 & -3 & 3 & -2 & 2 & -1 & 1 \\ -3 & 3 & -1 & 1 & -1 & 1 & -2 & -2 & 2 & 2 \\ -1 & 1 & -2 & 2 & -1 & 1 & 2 & -2 & 3 & -3 \\ -2 & 2 & -1 & 1 & -3 & 3 & -2 & 2 & -1 & 1 \\ -3 & 3 & -1 & 1 & -1 & 1 & -2 & -2 & 2 & 2 \\ -1 & 1 & -1 & 1 & -2 & 2 & -2 & -2 & 3 & -3 \\ -1 & 1 & -1 & -1 & 1 & 1 & +2 & -2 & 3 & -1 \\ -2 & 2 & -1 & 1 & -3 & 3 & -2 & 2 & -1 & 1 \\ -1 & 1 & -1 & -1 & 1 & 1 & +2 & -2 & 3 & -1 \end{bmatrix}$$

$$X$$

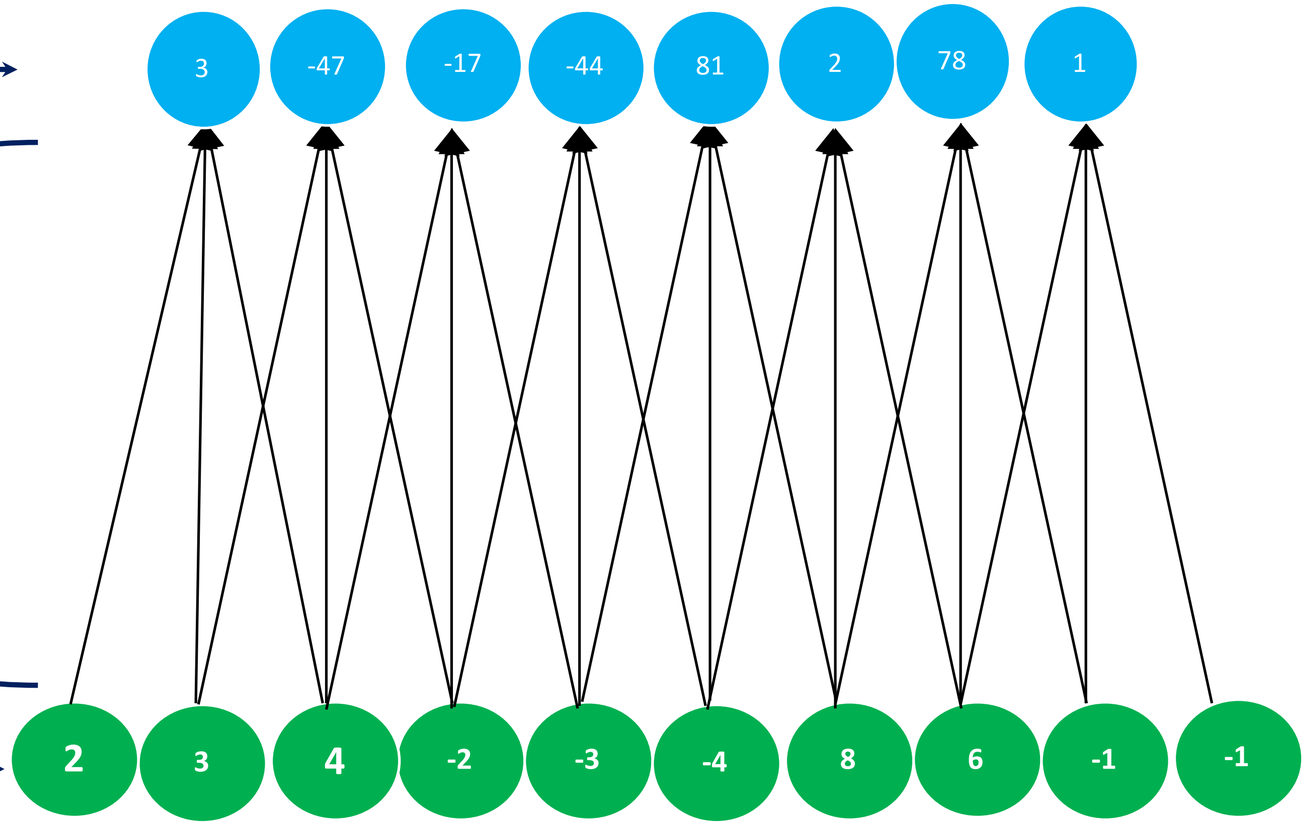


Dense connections

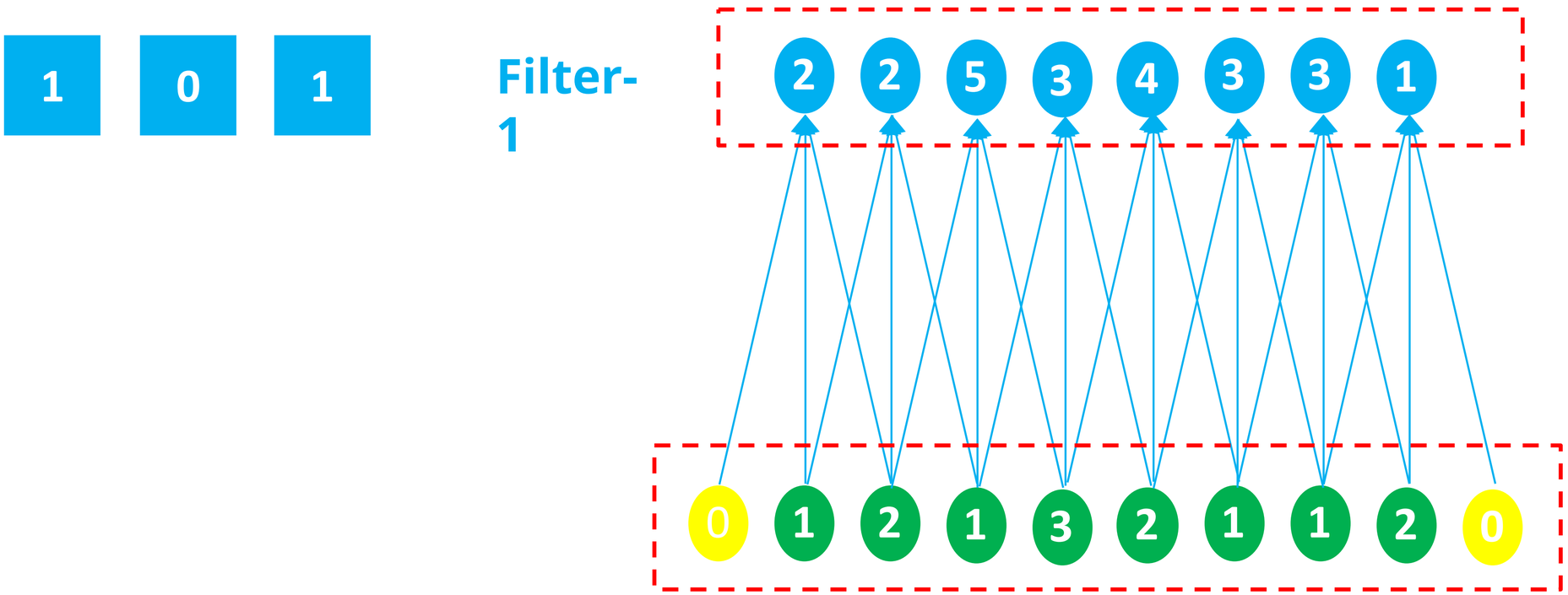
$$Y = W^T X$$

$$W = \begin{bmatrix} 6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -7 & 0 & 0 & 0 & 0 & 0 & 0 \\ -3 & -4 & -9 & 0 & 0 & 0 & 0 & 0 \\ 0 & 5 & 4 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & -9 & 0 & 5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 9 & -8 & 5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 8 & -4 & 9 & 0 \\ 0 & 0 & 0 & 0 & 0 & 9 & 0 & -2 \\ 0 & 0 & 0 & 0 & 0 & 0 & -6 & -7 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -6 \end{bmatrix}$$

$$X$$

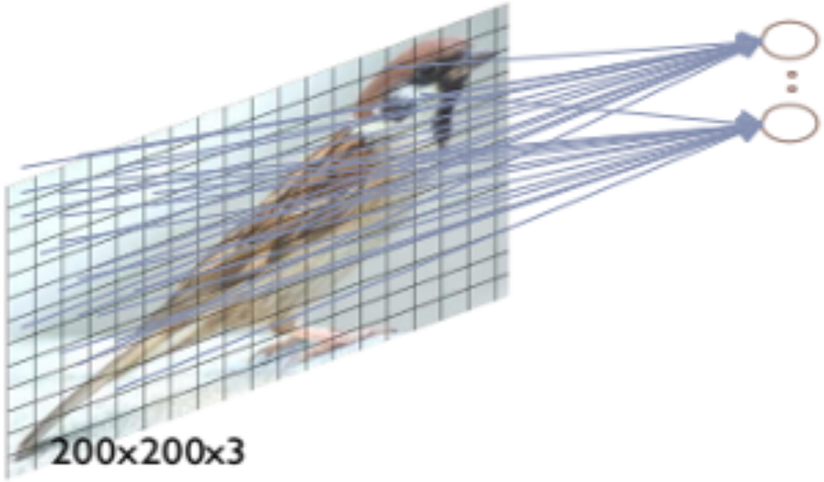


Convolutional connection



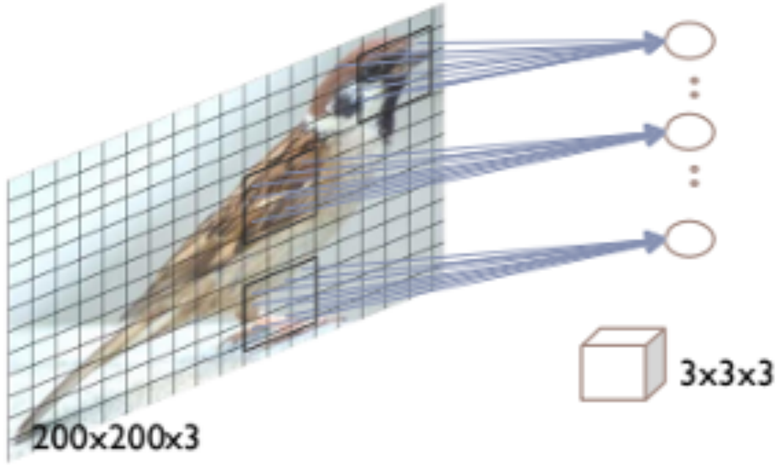
Convolution Layer

- Fully connected layer



- Image of size 200 X 200 and 3 colours (RGB)
- #Hidden Units: 120,000 (= 200X200X3)
- #Params: 14.4 billion (= 120K X 120K)
- Need huge training data to prevent over-fitting!

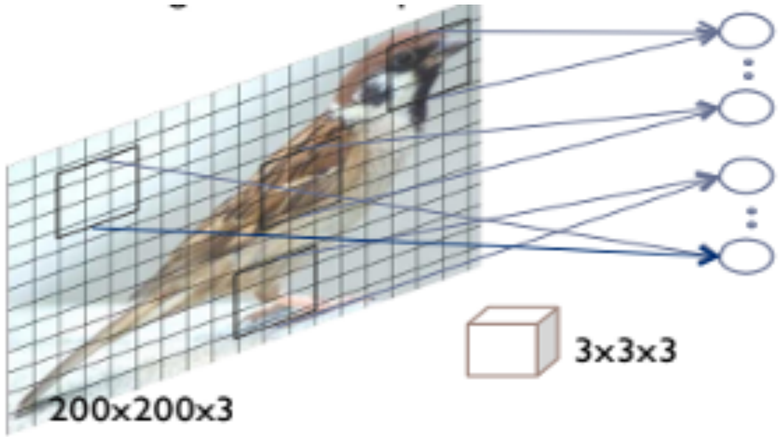
- Locally connected layer



- #Hidden Units: 120,000
- #Params: 3.2 Million (= 120K X 27)
- Useful when the image is highly registered

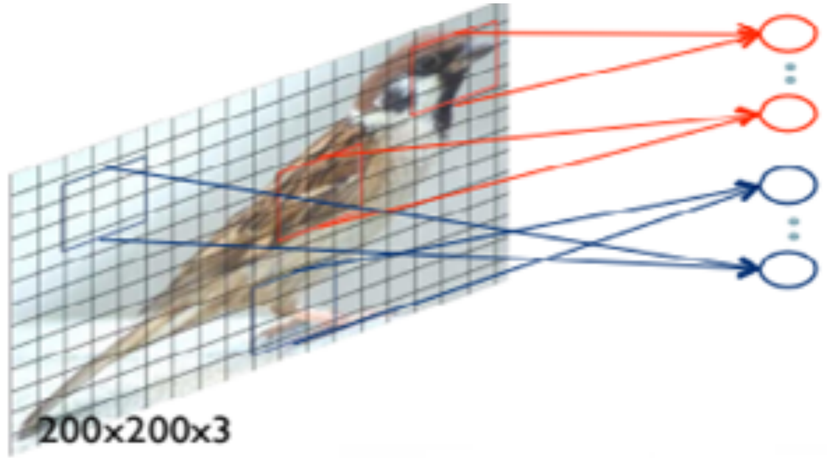
Convolution Layer

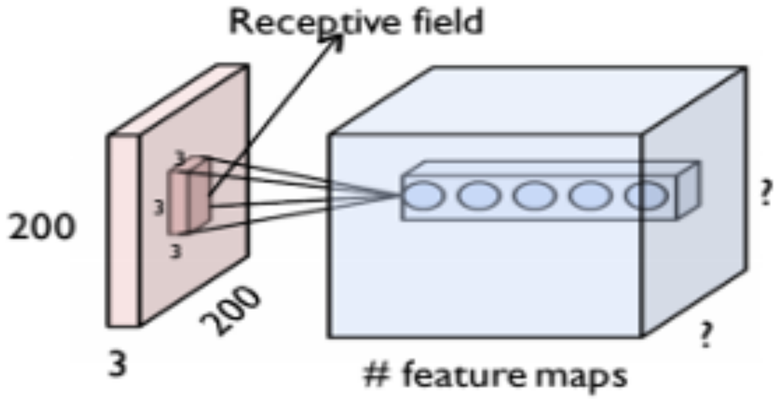
- Convolutional layer with single feature map.



- #Hidden Units: 120,000
- #Params: 27 x #Feature Maps
- Sharing parameters
- Exploiting the stationarity property and preserves locality of pixel dependencies

- Convolutional layer with multiple feature maps

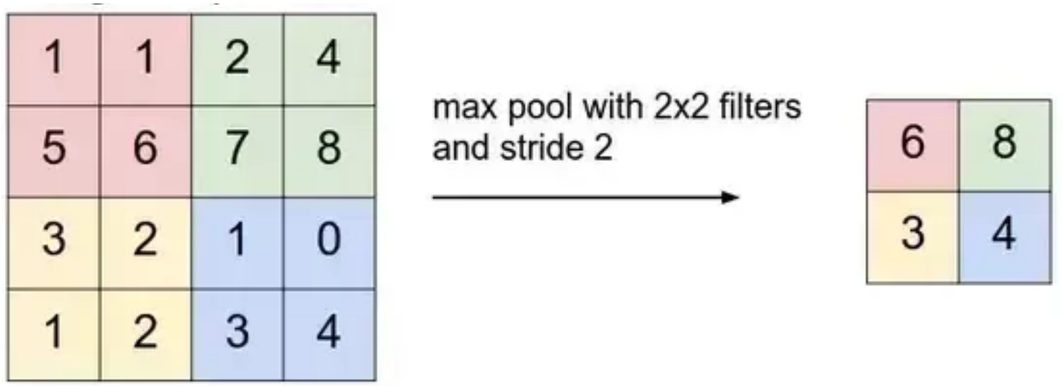




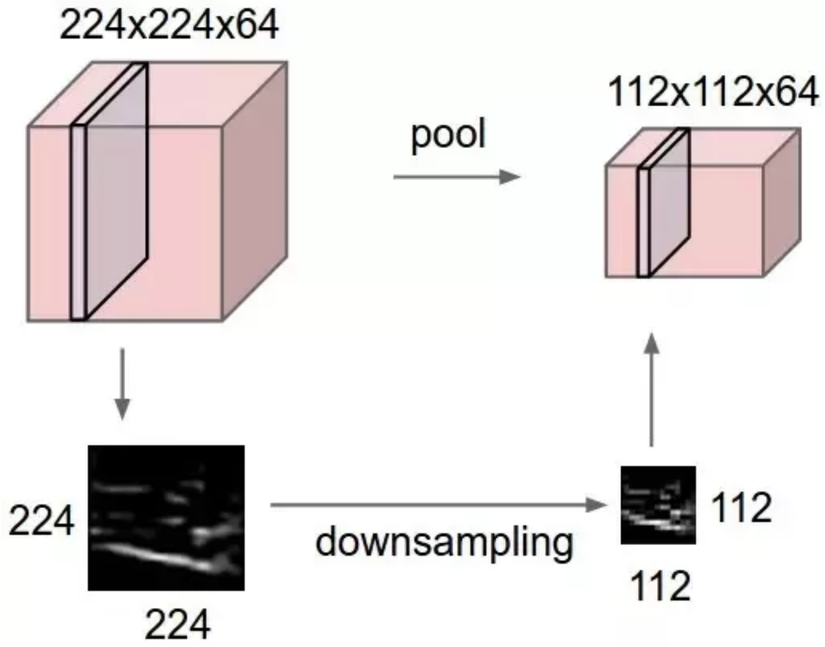
Stride and padding

For instance, here is a 3×3 kernel applied to a 5×5 input padded with a 1×1 border of zeros using 2×2 strides:

Pooling

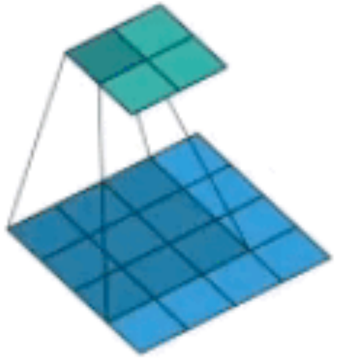
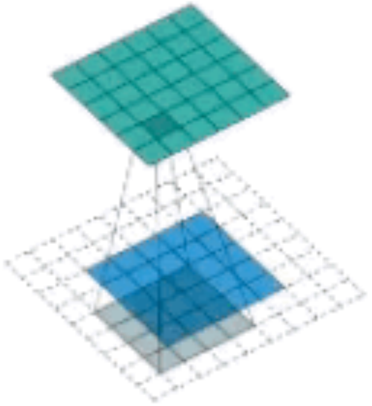
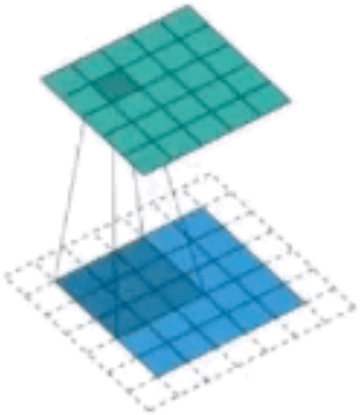
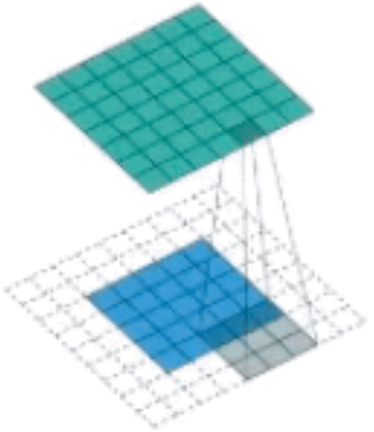
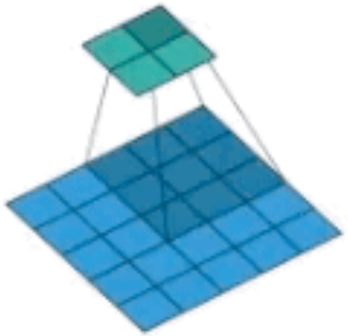
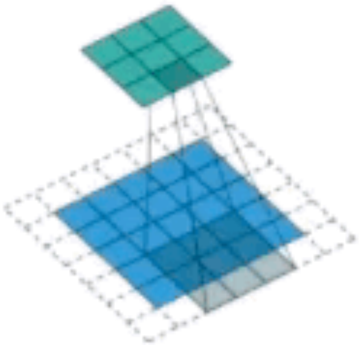
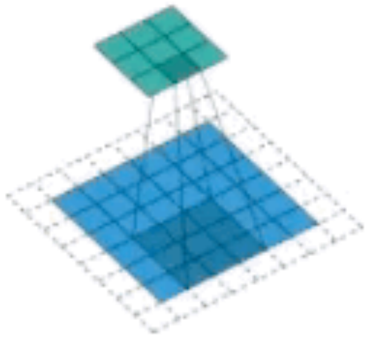


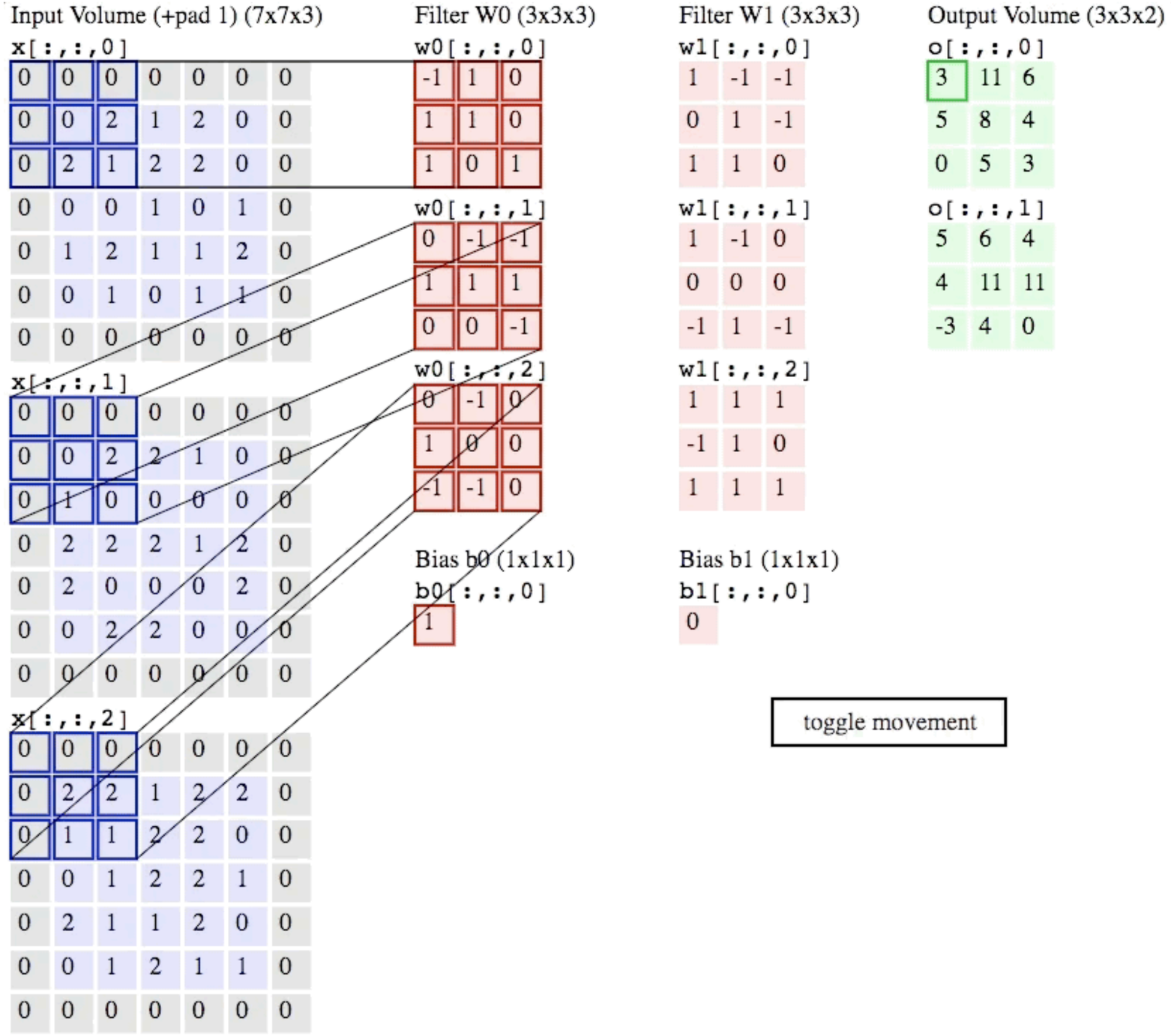
- **Motivation: We care about presence of features, not their exact location !**
- Dimensionality Reduction
- Prevents overfitting



Convolution animations

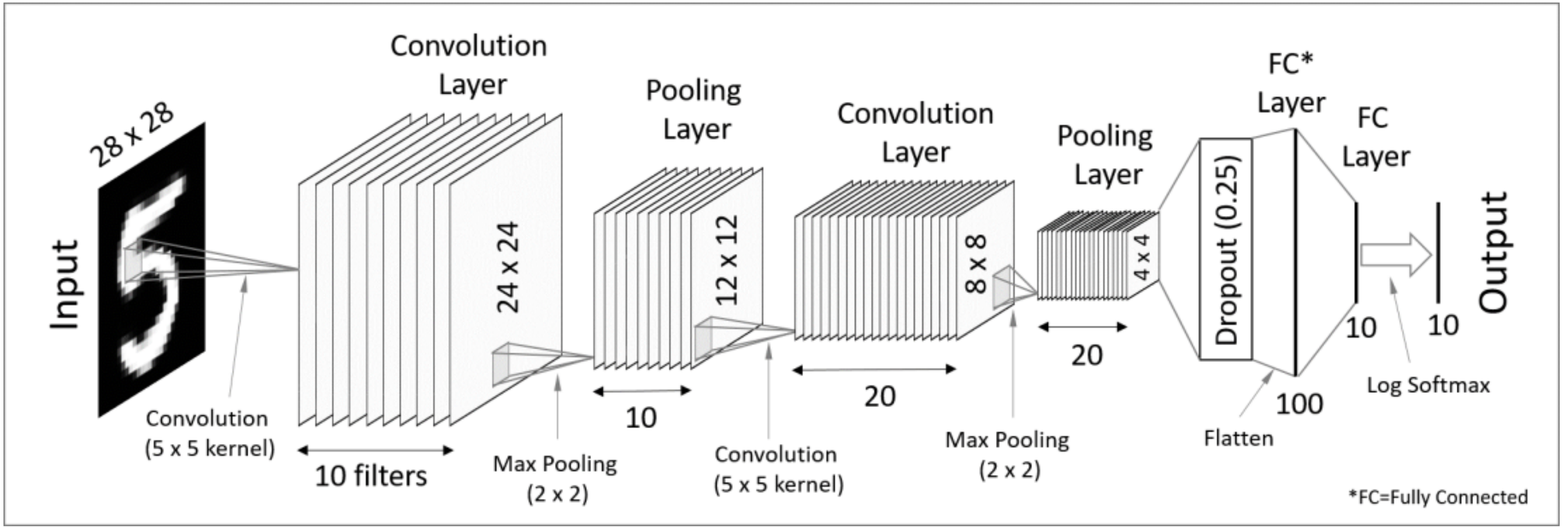
N.B.: Blue maps are inputs, and cyan maps are outputs.

			
No padding, no strides	Arbitrary padding, no strides	Half padding, no strides	Full padding, no strides
			
No padding, strides	Padding, strides	Padding, strides (odd)	

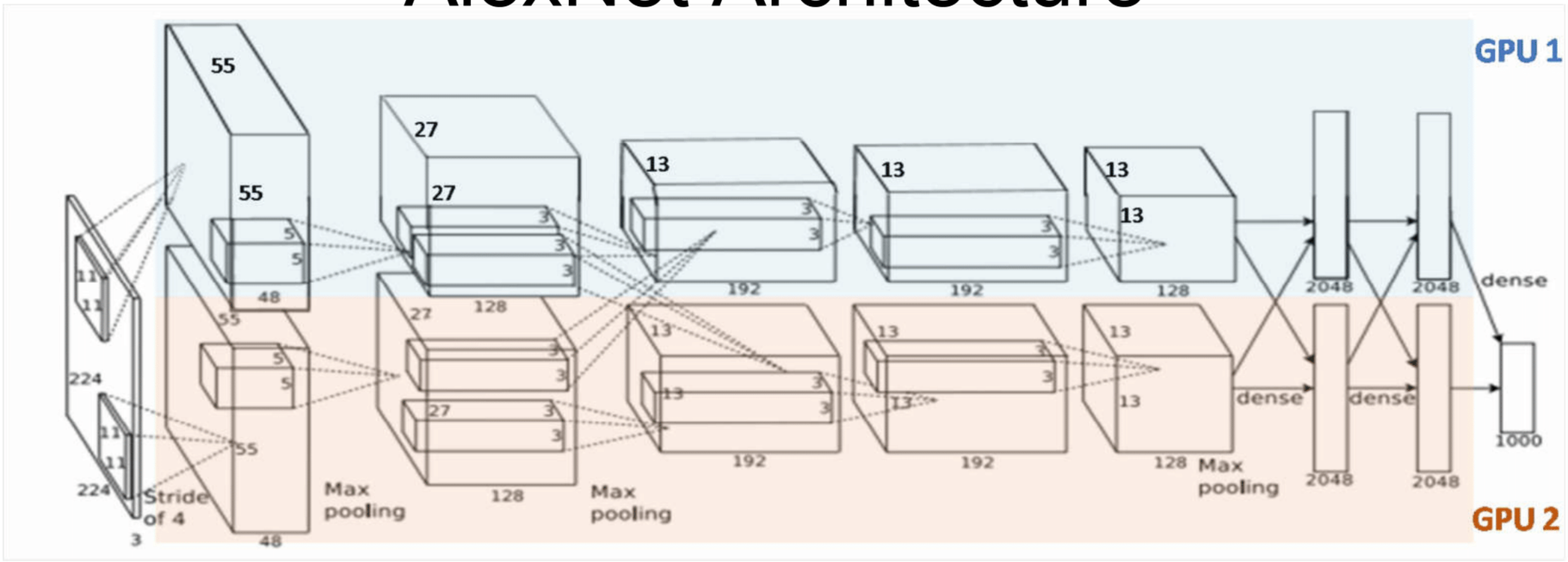

Input Volume (+pad 1) (7x7x3)
x[:,:,0]
0 0 0 0 0 0 0
0 0 2 1 2 0 0
0 2 1 2 2 0 0
0 0 0 1 0 1 0
0 1 2 1 1 2 0
0 0 1 0 1 1 0
0 0 0 0 0 0 0
x[:,:,1]
0 0 0 0 0 0 0
0 0 2 2 1 0 0
0 1 0 0 0 0 0
0 2 2 2 1 2 0
0 2 0 0 0 2 0
0 0 2 2 0 0 0
0 0 0 0 0 0 0
x[:,:,2]
0 0 0 0 0 0 0
0 2 2 1 2 2 0
0 1 1 2 2 0 0
0 0 1 2 2 1 0
0 2 1 1 2 0 0
0 0 1 2 1 1 0
0 0 0 0 0 0 0
Filter W0 (3x3x3)
w0[:,:,0]
-1 1 0
1 1 0
1 0 1
w0[:,:,1]
0 -1 -1
1 1 1
0 0 -1
w0[:,:,2]
0 -1 0
1 0 0
-1 -1 0
Bias b0 (1x1x1)
b0[:,:,0]
1
Filter W1 (3x3x3)
w1[:,:,0]
1 -1 -1
0 1 -1
1 1 0
w1[:,:,1]
1 -1 0
0 0 0
-1 1 -1
w1[:,:,2]
1 1 1
-1 1 0
1 1 1
Bias b1 (1x1x1)
b1[:,:,0]
0
Output Volume (3x3x2)
o[:,:,0]
3 11 6
5 8 4
0 5 3
o[:,:,1]
5 6 4
4 11 11
-3 4 0
toggle movement

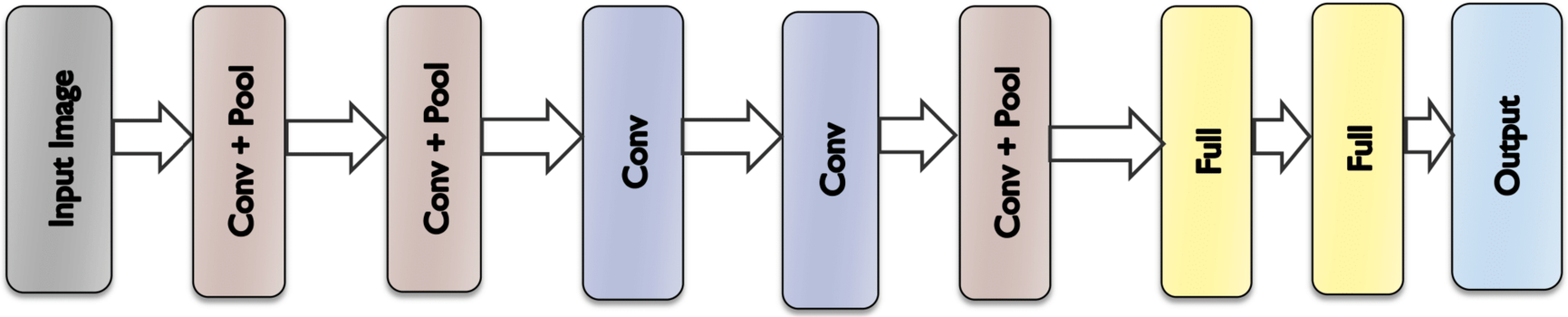
$$O = \frac{(W - K + 2P)}{S} + 1$$

where O is the output height/length, W is the input height/length, K is the filter size, P is the padding, and S is the stride.

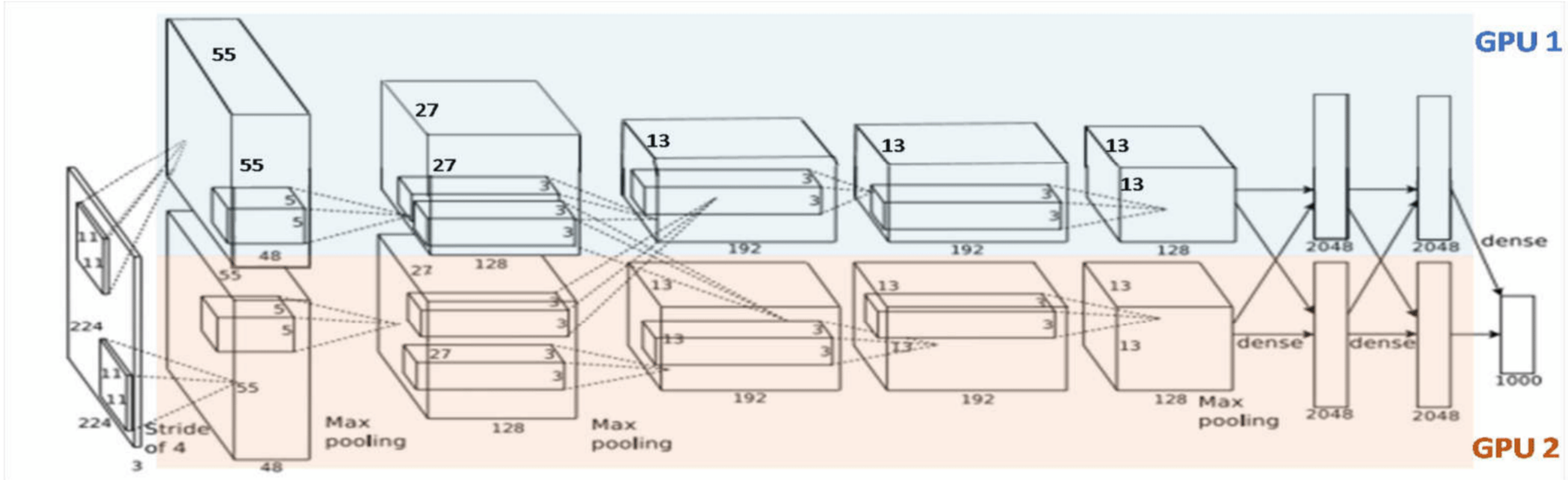

Input
28 x 28
Convolution
(5 x 5 kernel)
Convolution
Layer
10 filters
24 x 24
Max Pooling
(2 x 2)
Pooling
Layer
10
12 x 12
Convolution
(5 x 5 kernel)
Convolution
Layer
20
8 x 8
Max Pooling
(2 x 2)
Pooling
Layer
20
4 x 4
Flatten
Dropout (0.25)
100
FC*
Layer
10
FC
Layer
10
Log Softmax
Output
*FC=Fully Connected

AlexNet Architecture





AlexNet Architecture

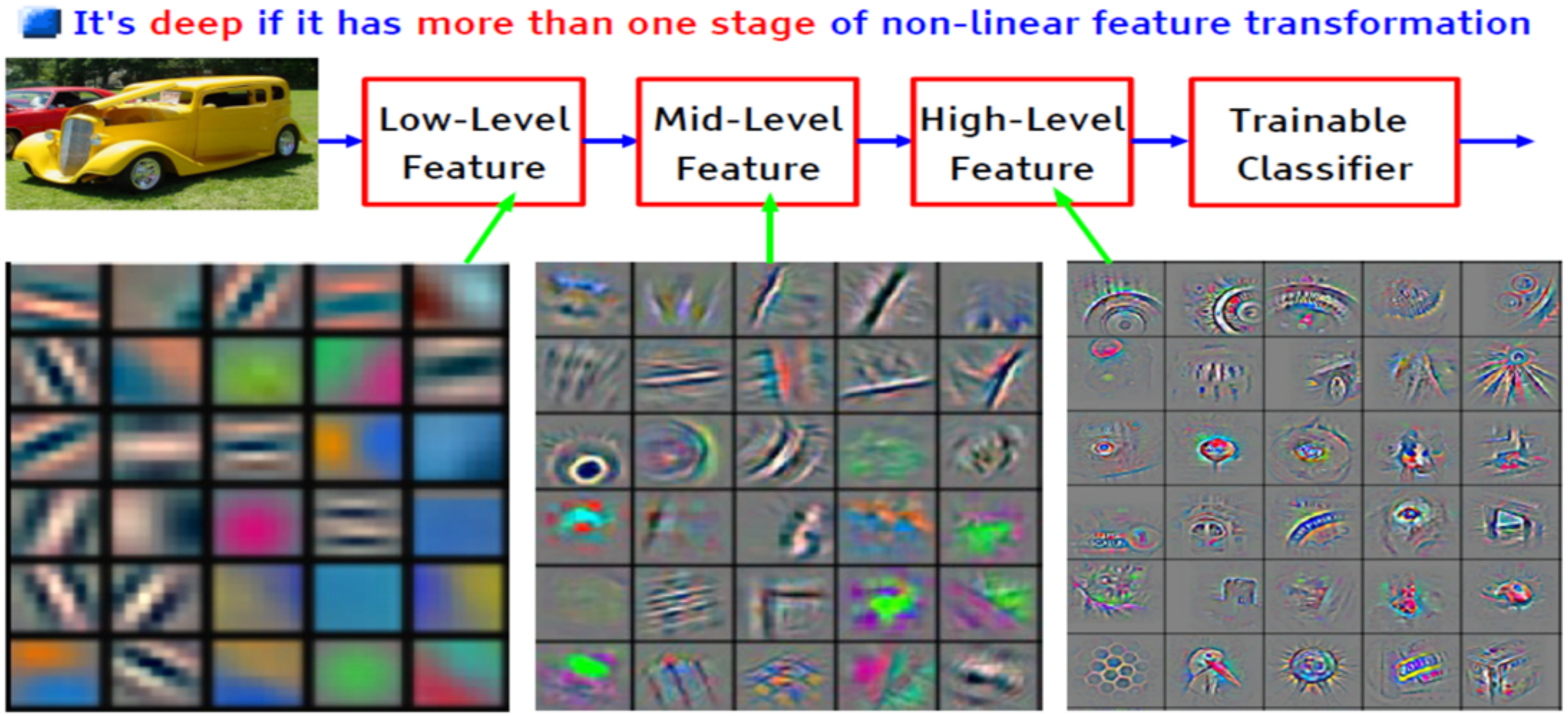


About 57 M parameters are in the fully connected layers

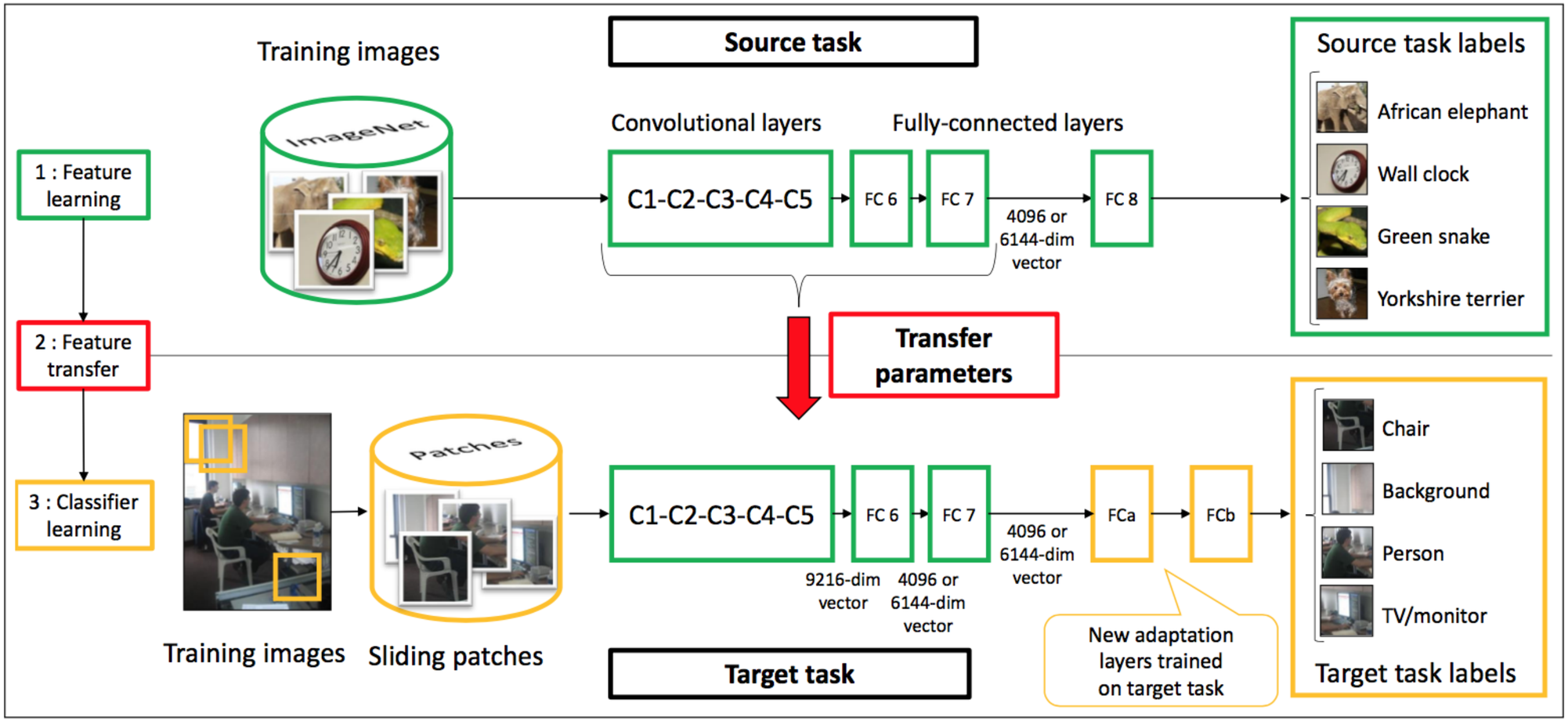
									Total
Parameters :	[(11 x 11 x 3) + 1] x 96 = 35 K	[5 x 5 x 48] x 256 = 307 K	[3 x 3 x 256] x 384 = 884 K	663 K	442 K	37 M	16 M	4 M	60 M
Neurons :	253,440	27 x 27 x 256 = 186,624	13 x 13 x 384 = 64,896	13 x 13 x 384 = 64,896	13 x 13 x 256 = 43,264	4096	4096	1000	0.63 M

- Convolutional layers cumulatively contain about 90-95% of computation, only about 5% of the parameters
- Fully-connected layers contain about 95% of parameters.

Deep Learnt Features

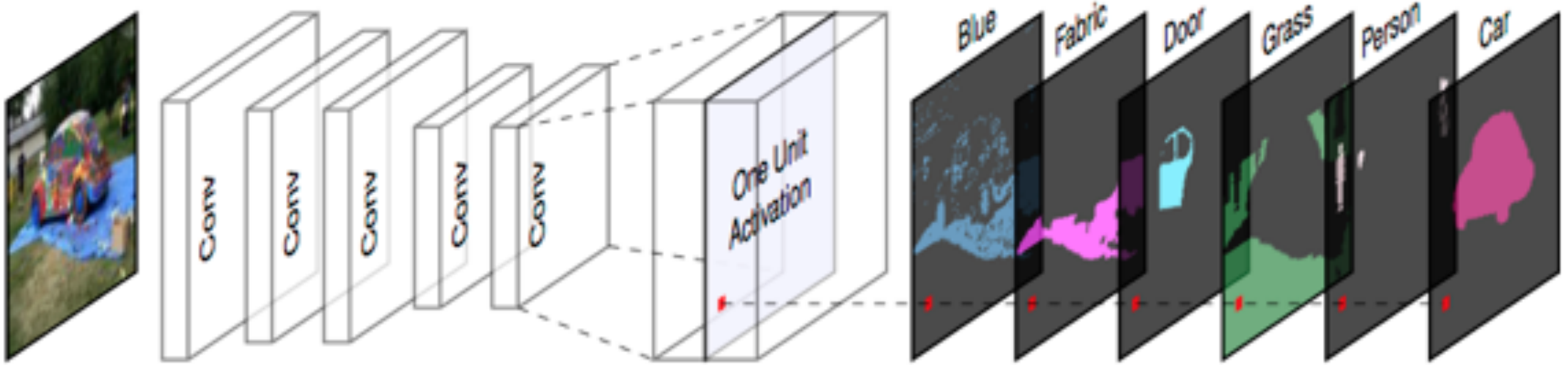


Transfer Learning



Input

Segmentation [9]


Conv
Conv
Conv
Conv
Conv
One Unit Activation
Blue
Fabric
Door
Grass
Person
Car