

Machine Learning: CNNs



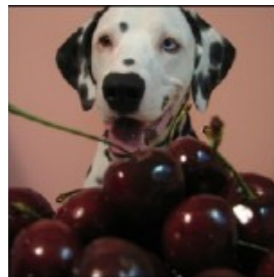
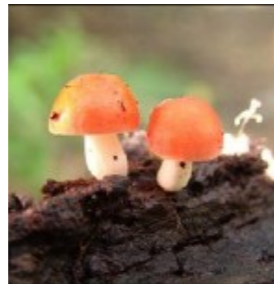
Outline

- ▶ Convolutional neural networks (CNNs)
 - why not use unstructured feed-forward models?
 - key parts: convolution, pooling
 - examples

Our problem: image classification

- ▶ E.g., image classification (1K categories)

Image



...

Category

mushroom

cherry

...



Feed-forward networks



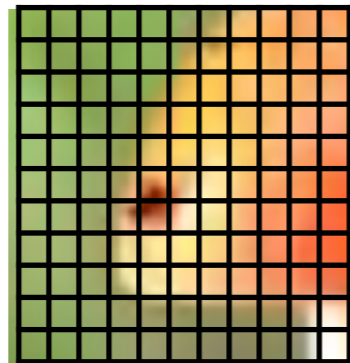
input



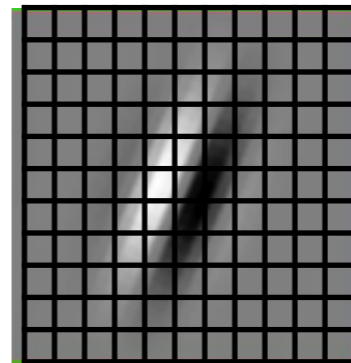
layer 1



Patch classifier/filter



11x11
input



11x11
weights

Convolution

$$f = \text{ReLU} \left(\begin{array}{c} \text{11x11} \\ \text{input} \end{array} \cdot \begin{array}{c} \text{11x11} \\ \text{weights} \end{array} \right)$$



input



feature map

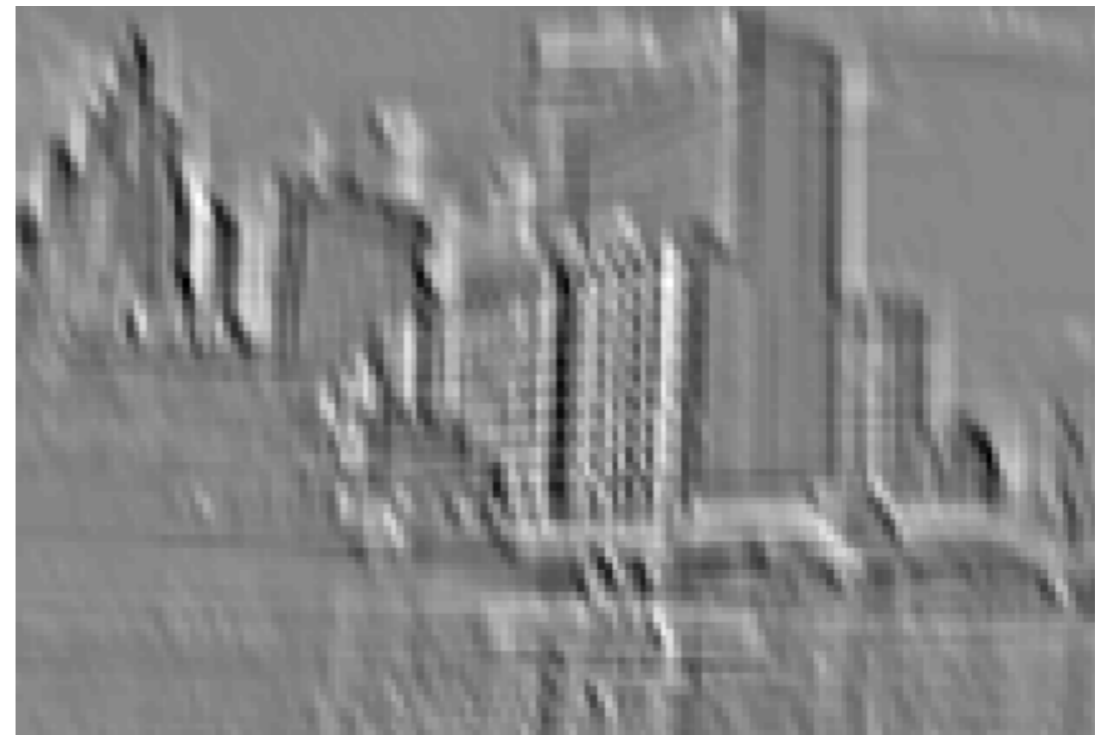


Convolution, feature map

filter
patch



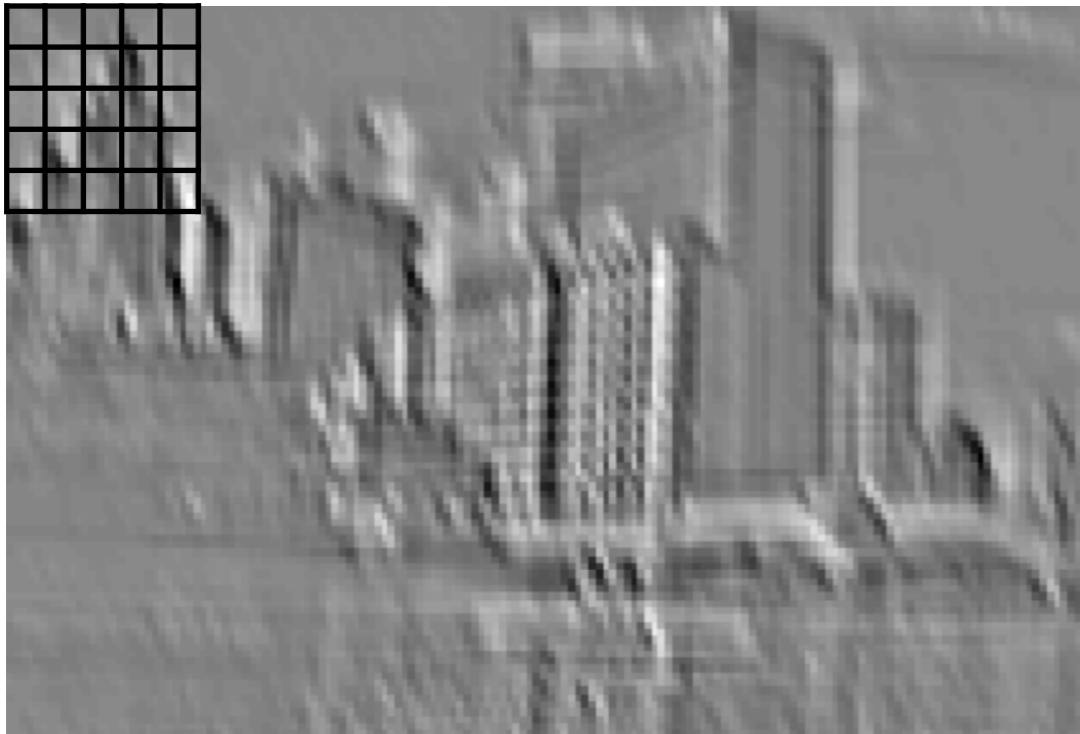
original image



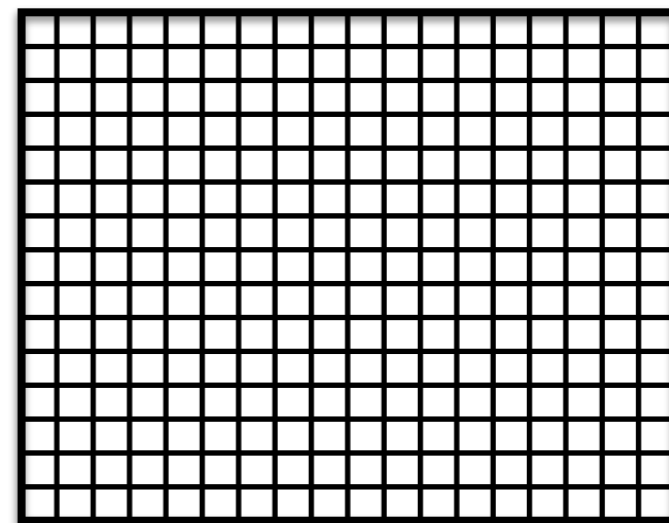
resulting feature map

Pooling

- ▶ We wish to know whether a feature was there but not exactly where it was



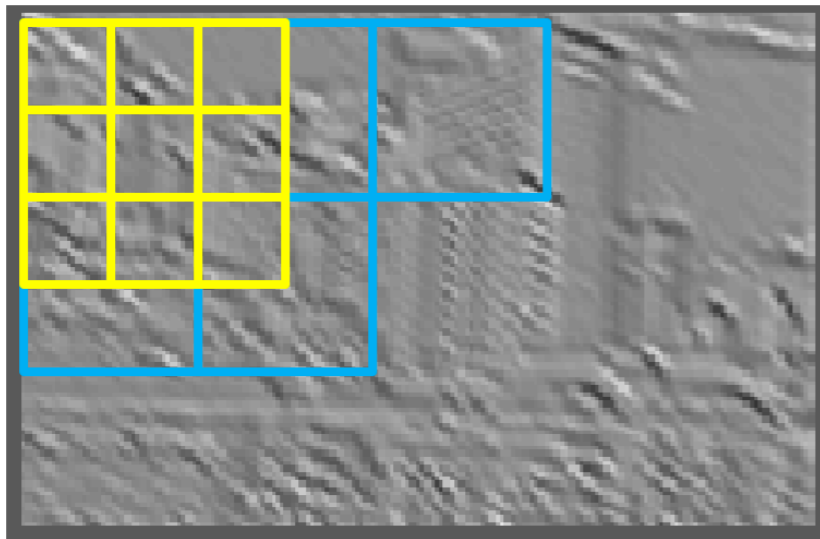
feature map



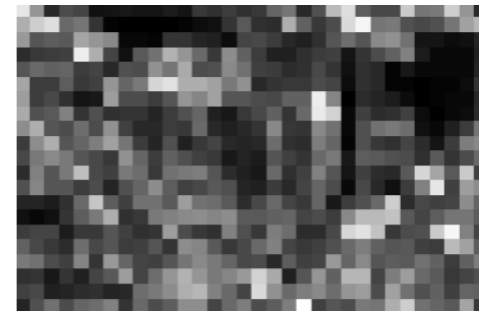
pooled map

Pooling (max)

- ▶ Pooling region and “stride” may vary
 - pooling induces translation invariance at the cost of spatial resolution
 - stride reduces the size of the resulting feature map

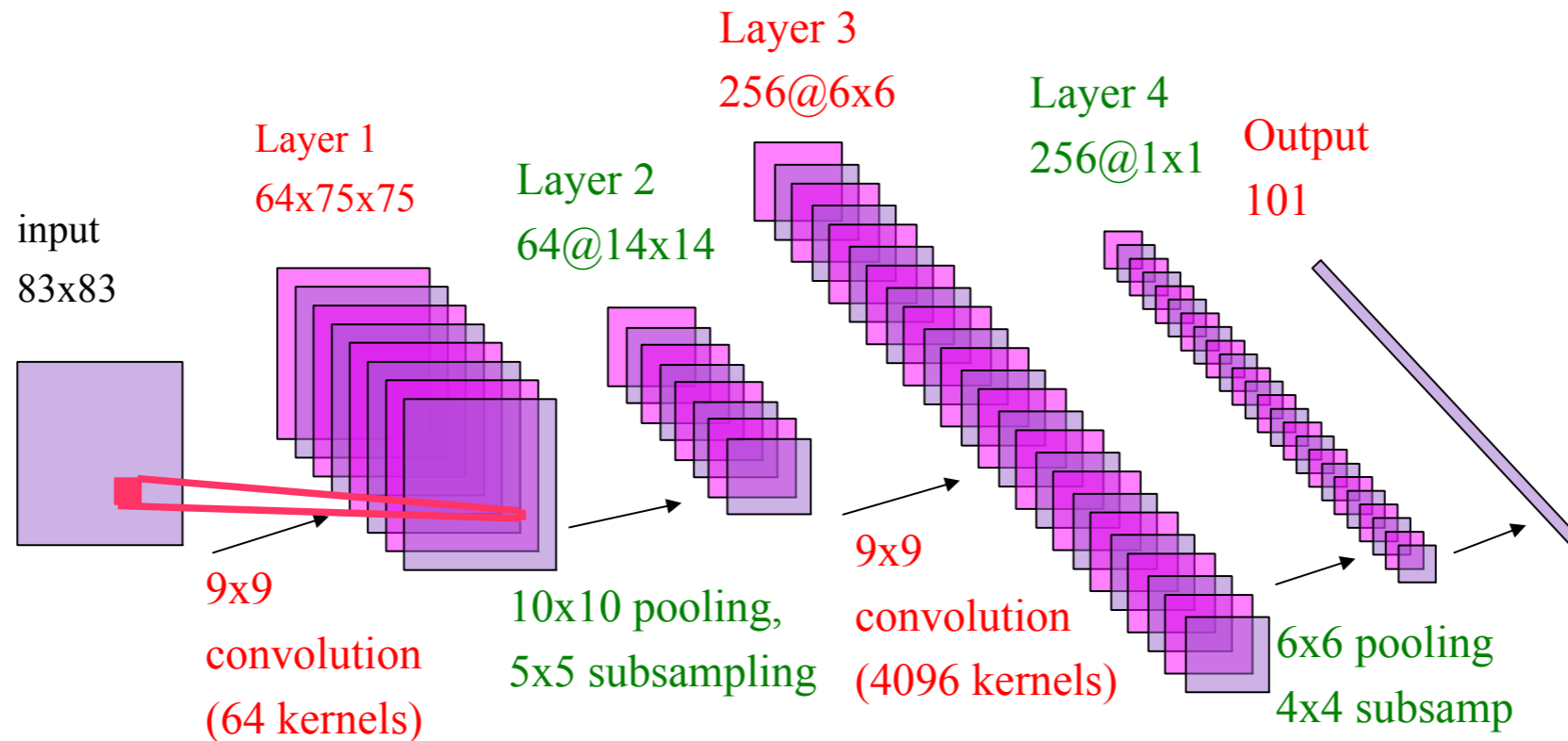


feature map



feature map
after max pooling

Convolutional Neural Network



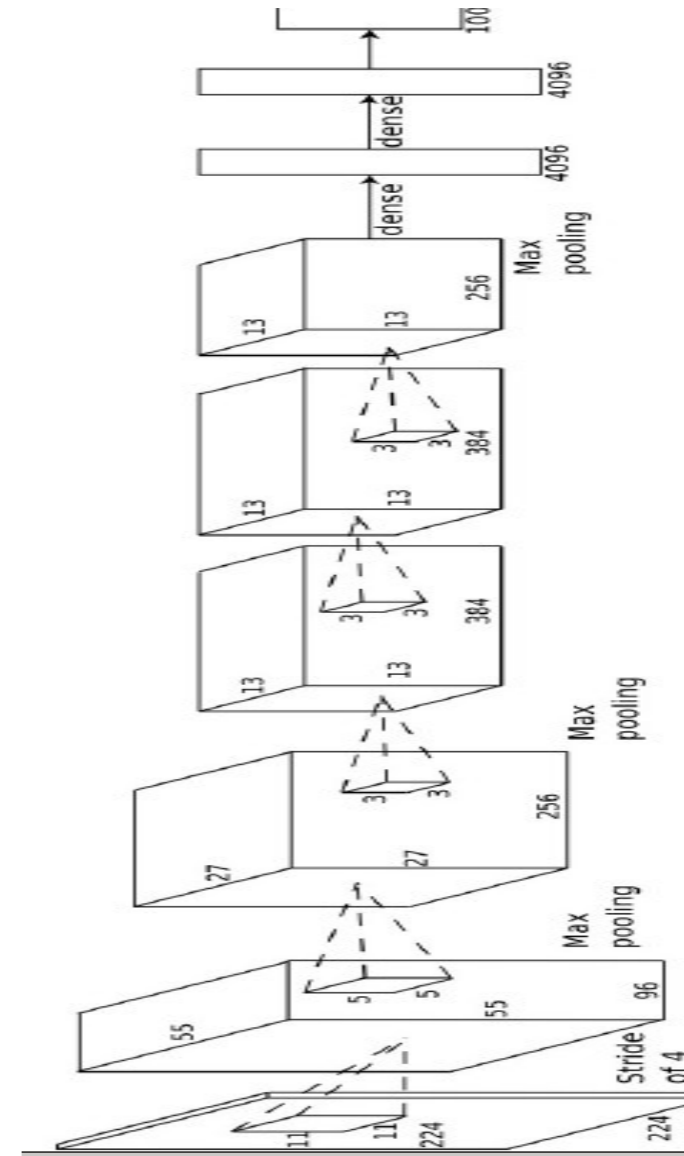
- **Non-Linearity:** half-wave rectification, shrinkage function, sigmoid
- **Pooling:** average, L1, L2, max
- **Training:** Supervised (1988-2006), Unsupervised+Supervised (2006-now)

(LeCun 13')

Convolutional Neural Network

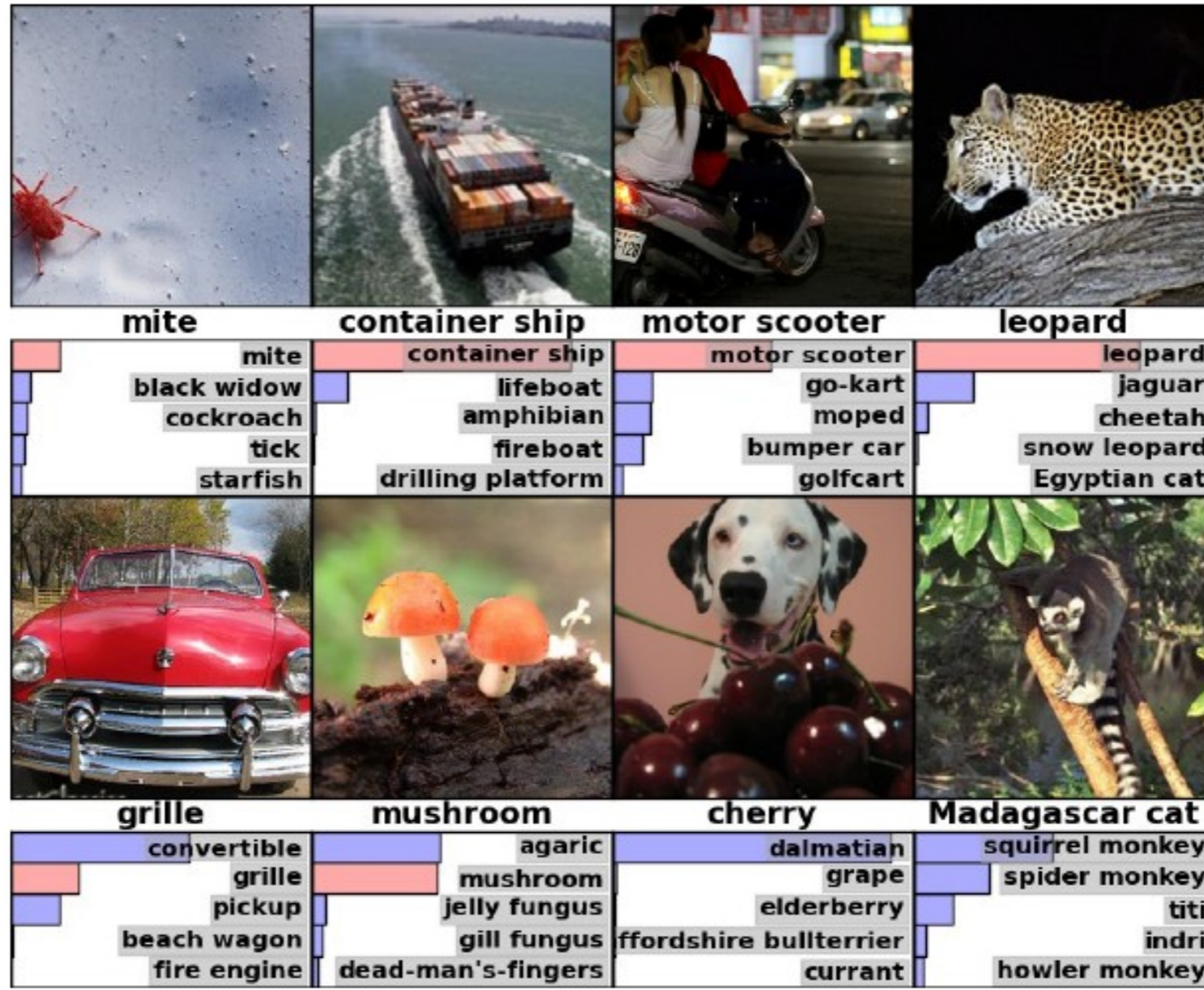
- 1000 categories, 1.5 Million labeled training samples
- Won the 2012 ImageNet LSVRC. 60 Million parameters, 832M MAC ops

4M	FULL CONNECT	4Mflop
16M	FULL 4096/ReLU	16M
37M	FULL 4096/ReLU	37M
	MAX POOLING	
442K	CONV 3x3/ReLU 256fm	74M
1.3M	CONV 3x3ReLU 384fm	224M
884K	CONV 3x3/ReLU 384fm	149M
	MAX POOLING 2x2sub	
	LOCAL CONTRAST NORM	
307K	CONV 11x11/ReLU 256fm	223M
	MAX POOL 2x2sub	
	LOCAL CONTRAST NORM	
35K	CONV 11x11/ReLU 96fm	105M



(Krizhevsky et al., 12')

Convolutional Neural Network



(Krizhevsky et al., 12')

ConvNet features

Learned layer 1 CNN filters



96 convolutional filters on the first layer
(filters are of size $11 \times 11 \times 3$, applied across
input images of size $224 \times 224 \times 3$)

(Krizhevsky et al., 12')