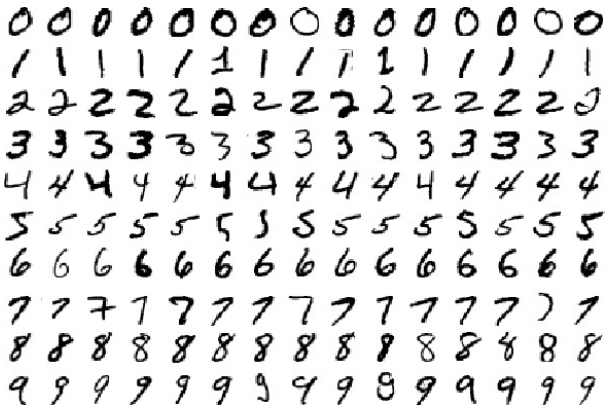
Convolutional Neural Net

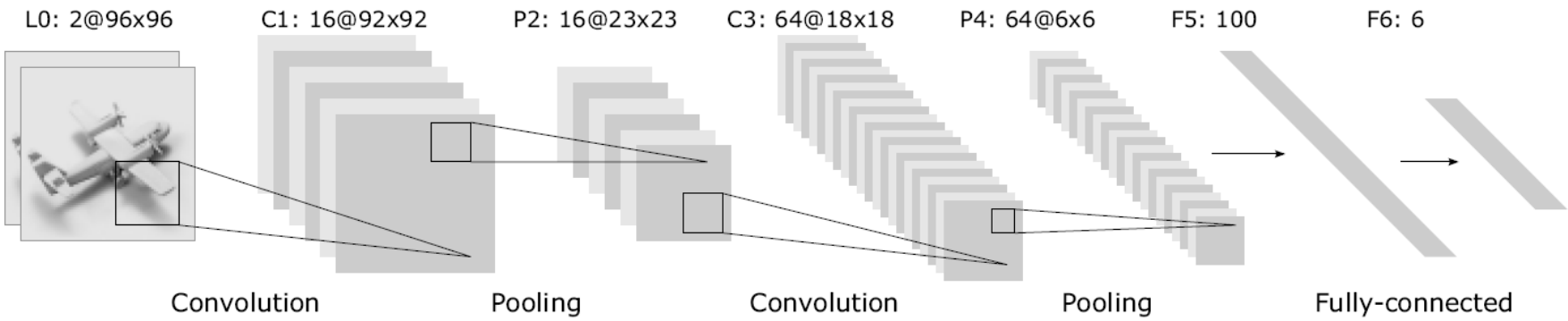
COMP4211

Handwritten Digit Recognition

MNIST: 10 classes (digits 0to 9)

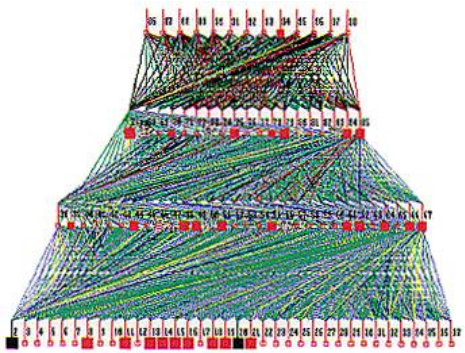


Convolutional neural network



Local Receptive Fields

- standard MLP



- local receptive fields
  - inspired from biology
  - cf. image processing

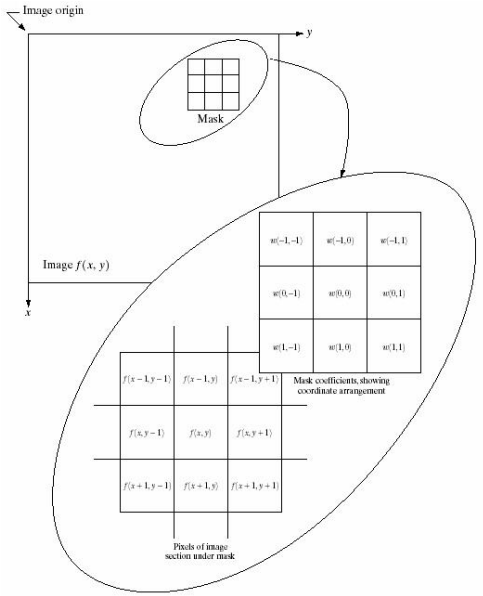


FIGURE 3.32 The mechanics of spatial filtering. The magnified drawing shows a 3×3 mask and the image section directly under it; the image section is shown displaced out from under the mask for ease of readability.

Mask (Convolution Mask)

z1	z2	z3
z4	z5	z6
z7	z8	z9

w1	w2	w3
w4	w5	w6
w7	w8	w9

- $f(x, y)$ is centered around z_5
- w_i: mask coefficient
- response of a linear mask: $\sum_{i=1}^{9} w_i z_i$

Smoothing (Averaging) Filter

$$1/9 * \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

- window size

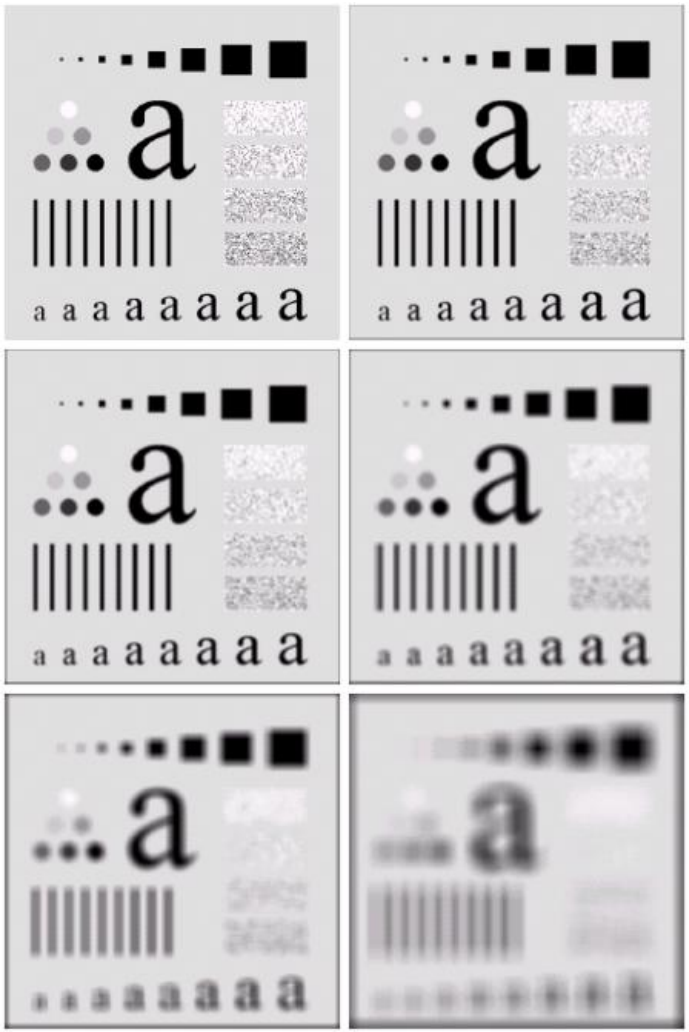
original

n=5 (nxn mask)

n=15 (nxn mask)

n=25 (nxn mask)

Examples



a b
c d
e f

FIGURE 3.35 (a) Original image, of size 500 × 500 pixels. (b)–(f) Results of smoothing with square averaging filter masks of sizes $n = 3, 5, 9, 15$, and 35, respectively. The black squares at the top are of sizes 3, 5, 9, 15, 25, 35, 45, and 55 pixels, respectively; their borders are 25 pixels apart. The letters at the bottom range in size from 10 to 24 points, in increments of 2 points; the large letter at the top is 60 points. The vertical bars are 5 pixels wide and 100 pixels high; their separation is 20 pixels. The diameter of the circles is 25 pixels, and their borders are 15 pixels apart; their gray levels range from 0% to 100% black in increments of 20%. The background of the image is 10% black. The noisy rectangles are of size 50 × 120 pixels.

Other Arrangements

1	1	1
1	2	1
1	1	1

1	1	1	1	1
1	2	3	2	1
1	3	4	3	1
1	2	3	2	1
1	1	1	1	1

center pixel: 1 vs 5

Sharpening Filters

Averaging pixels

- blur
- analogous to integration, related to sum of pixel intensity values

Differentiation

- has the opposite effect of blurring
- sharpens an image, related to difference between intensity values

First derivative

$$\frac{\partial f}{\partial x} \leftrightarrow f(x+1) - f(x)$$

First Order Derivatives

Gradient: a vector

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \\ \frac{\partial f}{\partial y} \end{bmatrix}$$

- for each (x, y) you are storing two values
- often have two images to represent this: X-gradient and Y-gradient (can be computed independently)

Example: X-Gradient and Y-Gradient

original

df/dx

df/dy

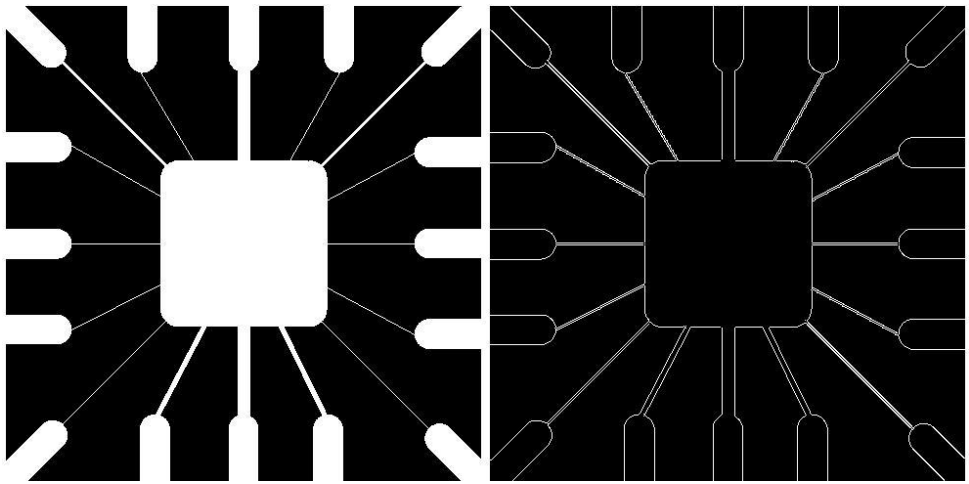
$\text{mag}(\nabla f)$

Edge Detector

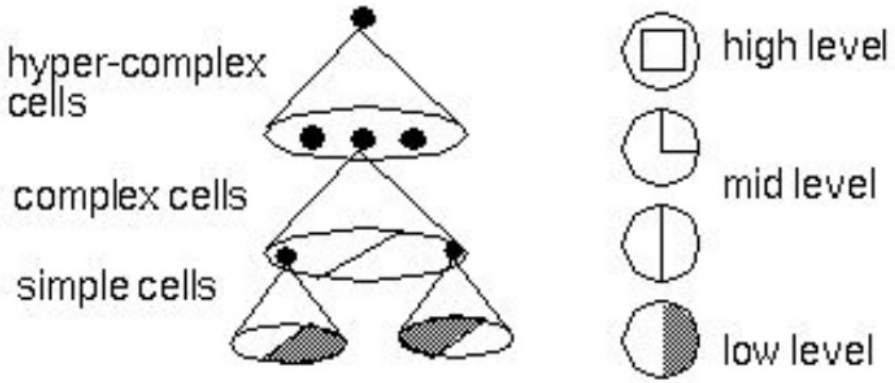
-1	0	1
-1	0	1
-1	0	1

-1	-1	-1
0	0	0
1	1	1

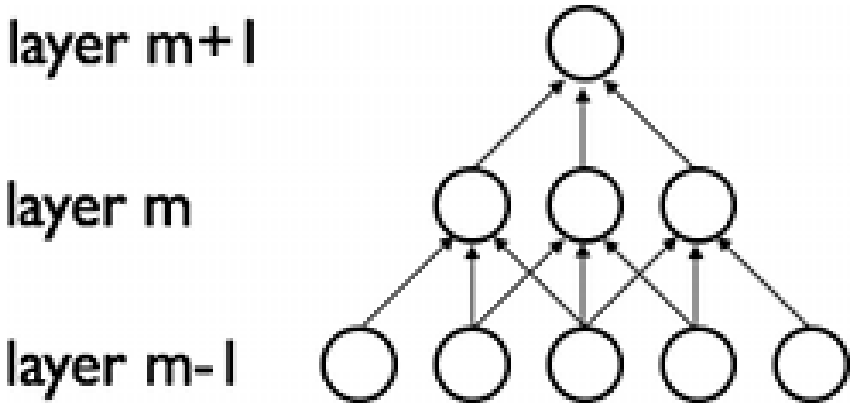
- difference between first and third column (df/dy)
- difference between first and third row (df/dx)



Feature Hierarchy

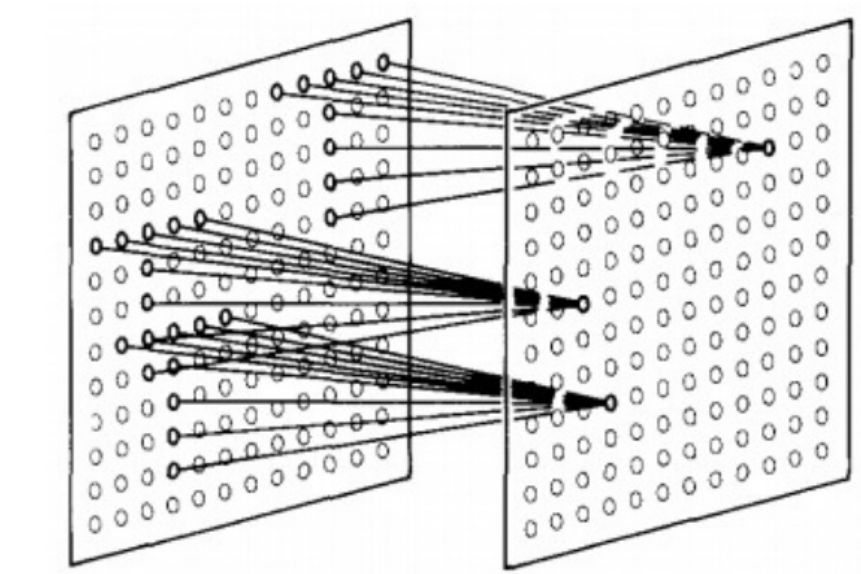


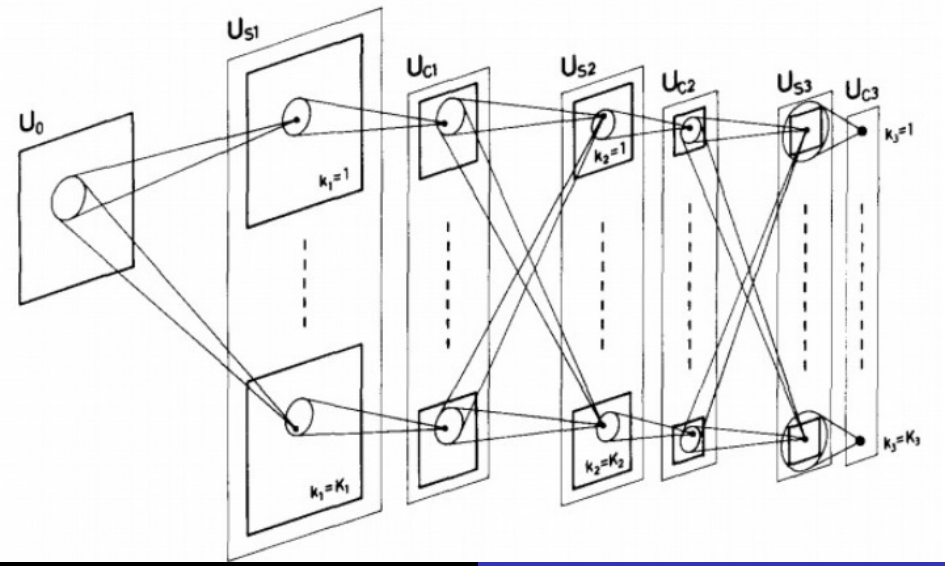
- hidden units are connected to a local subset of units in the previous layer



Feature Hierarchy...

- another early model: Neurocognitron [Fukushima 1980]





Shared Weights

- each local receptive field is replicated across the entire image

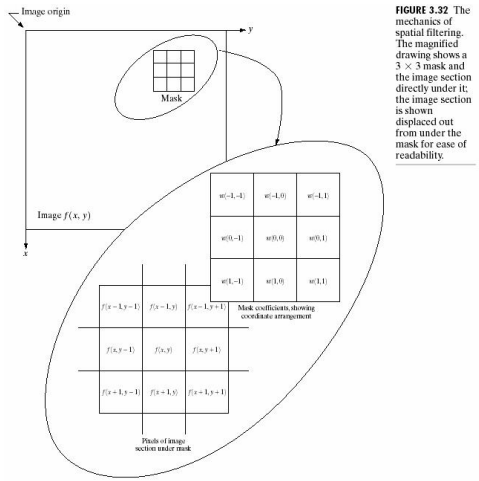
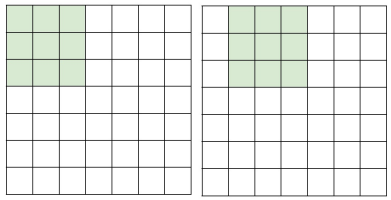
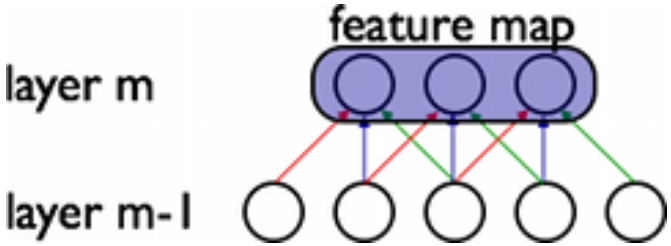


FIGURE 3.32 The mechanics of spatial filtering. The magnified drawing shows a 3×3 mask and the image section directly under it; the image section is shown displaced out from under the mask for ease of readability.

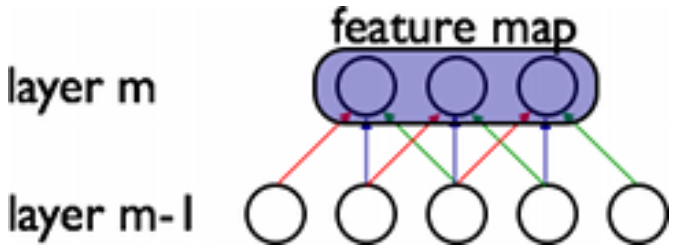


  - weights of the same color are shared (constrained to be identical)

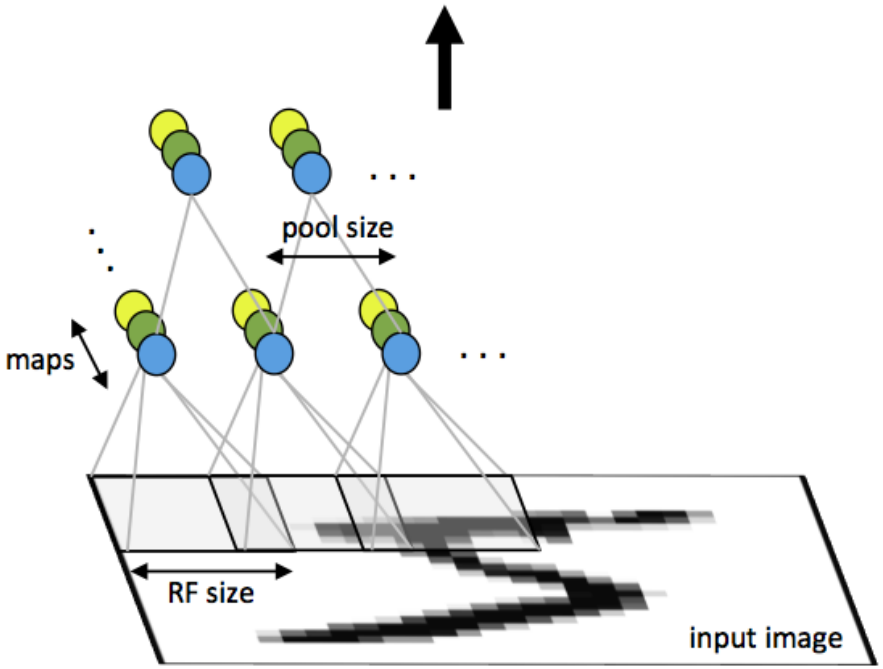


- allows for features to be detected regardless of their position in the image
  - robustness to shifts of the input
- greatly reduces the number of free parameters to learn

Convolutional Layer



- multiple feature maps look at the same region of the input



Pooling Layer

motivation: spatial invariance

- once a feature has been detected, only its approximate position relative to other features is relevant

Example

the input image contains

1. the endpoint of a roughly horizontal segment in the upper left area
2. a corner in the upper right area
3. the endpoint of a roughly vertical segment in the lower portion

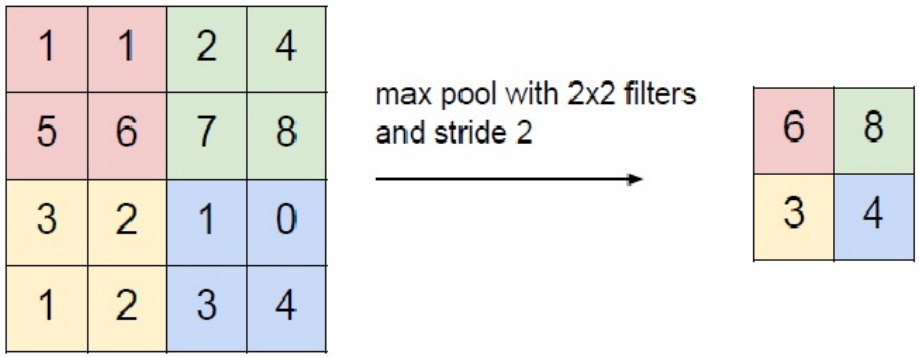
the input image is a seven

- positions are likely to vary for different instances of the character

Spatial Downsampling

max-pooling

- for each such sub-region (e.g., over a 2×2 area in the previous layer), outputs the maximum value



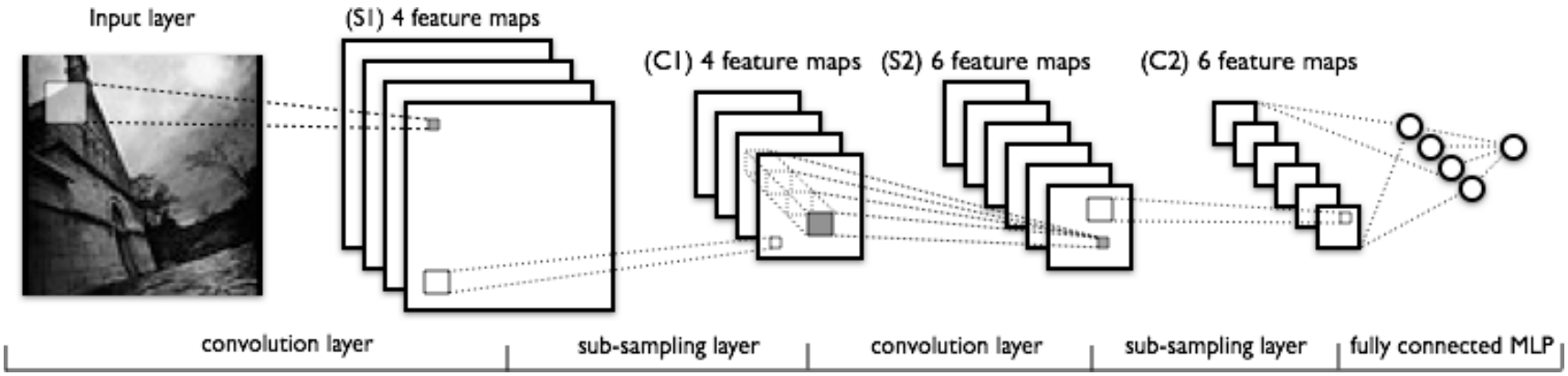
subsampling

- local averaging
- multiplies by a trainable coefficient, adds a trainable bias, and passes the result through a sigmoid function

partitions the input image into non-overlapping rectangles

- reduces the resolution of the feature map
- e.g., half the number of rows and columns as the feature maps in the previous layer

Example



- lower-layers: alternating convolution and max-pooling layers
- fully-connected (traditional MLP)
- classification error

Application: Face Recognition