Convolutional Neural Network

Introduction

Why AI in Computer Vision?

1. one picture worth a thousand words
2. large video and image collections

How many images and videos are there on the internet?

Social media statistics

- The internet has 4.2 billion users
- 3.03 billion active social media users
- Youtube: 300 hours of video are uploaded every minute
- Instagram: Over 95 million photos are uploaded to each day; More than 40 billion photos have been shared so far

typical computer vision tasks

Object Recognition

- identify objects and scenes

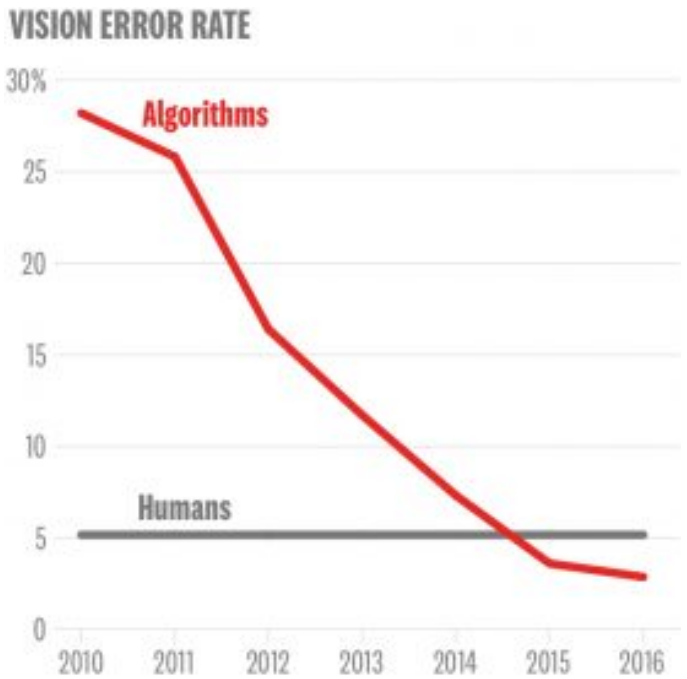
Error Rate

ImageNet Challenge

IMAGENET

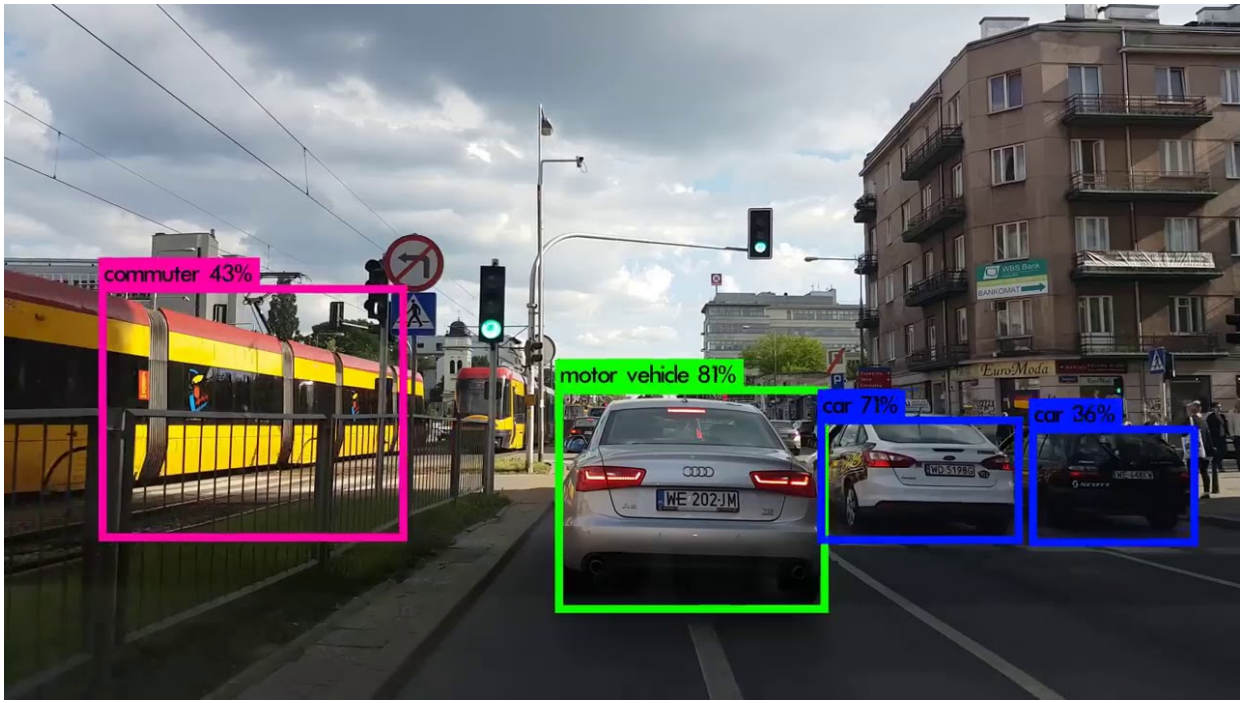
- 1,000 object classes (categories).
- Images:
  - 1.2 M train
  - 100k test.

4



Real-Time Object Detection

YOLO 9000: Can detect over 9000 object categories



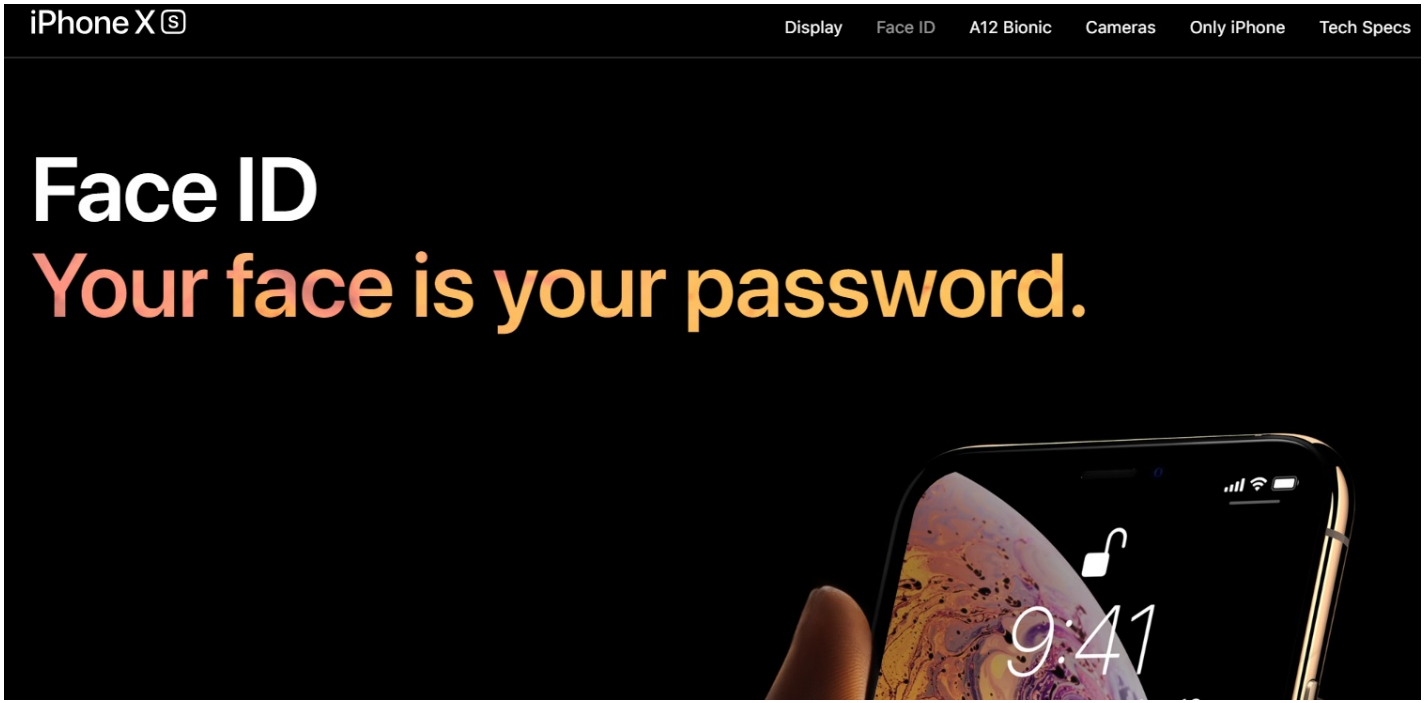
(video)

Example: Self-Driving Trucks

One of 10 Breakthrough Technologies in 2017

(video)

Face Recognition



Example: Paying with Your Face

One of 10 Breakthrough Technologies in 2017

(link)

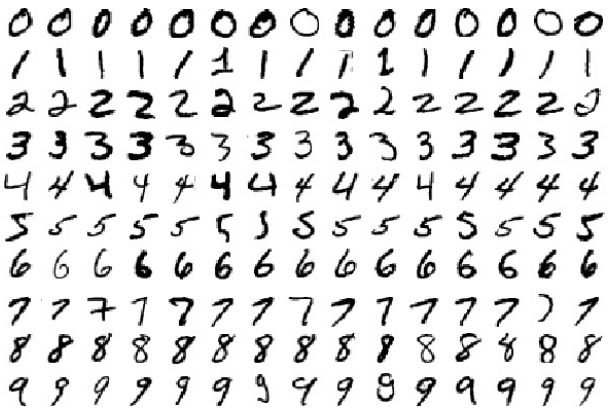
More Generally, Biometrics

(link)

Key AI technology:
Convolutional neural network (CNN)

Handwritten Digit Recognition

MNIST: 10 classes (digits 0to 9)



Convolutional neural network

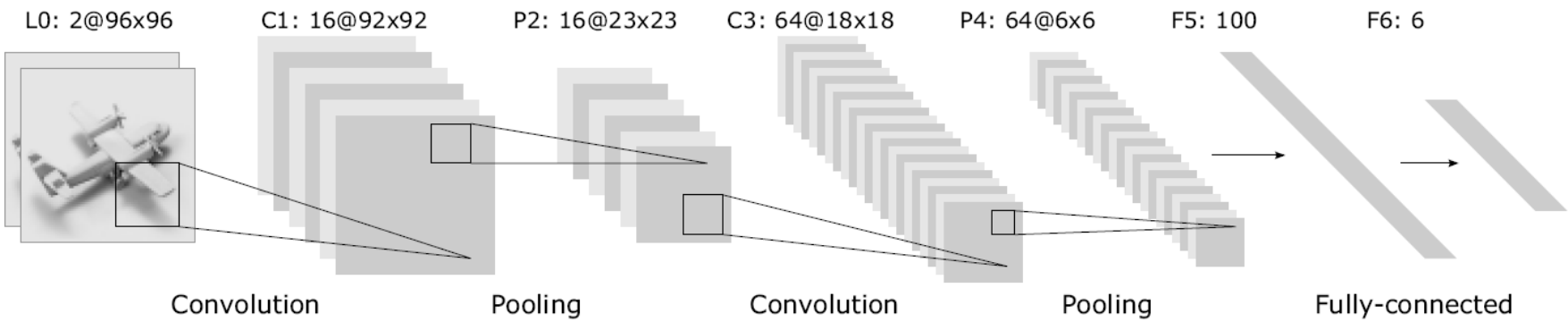


Image Processing Basics: 2D Convolution

image $I(i, j)$; kernel (mask) K

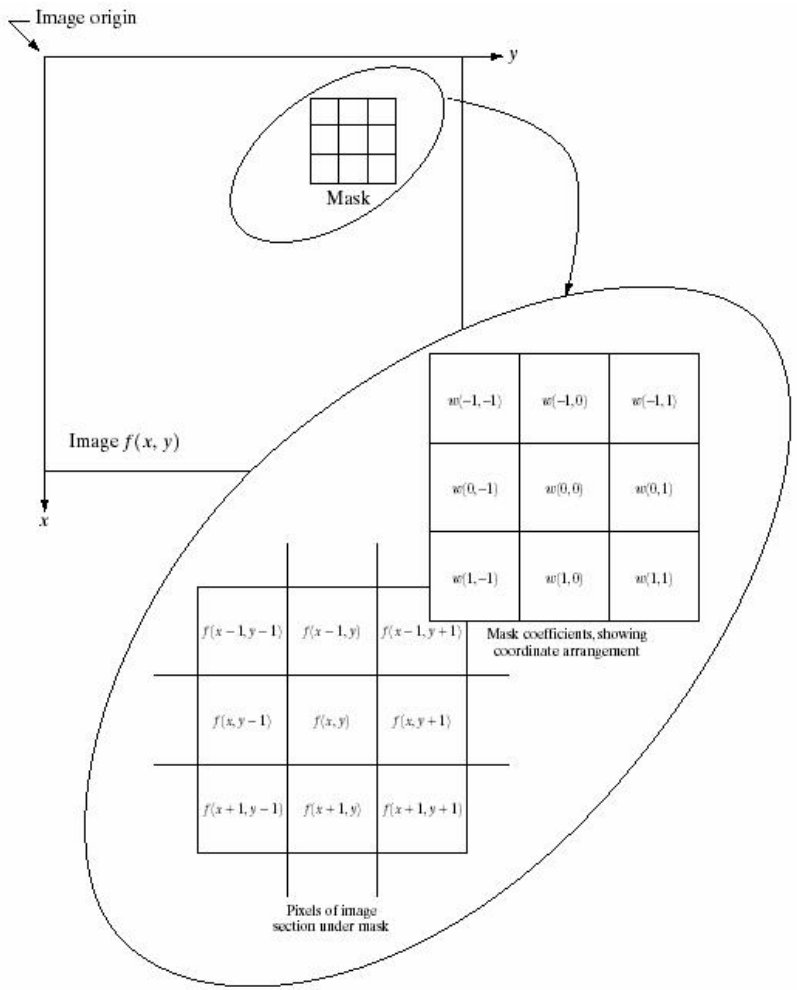
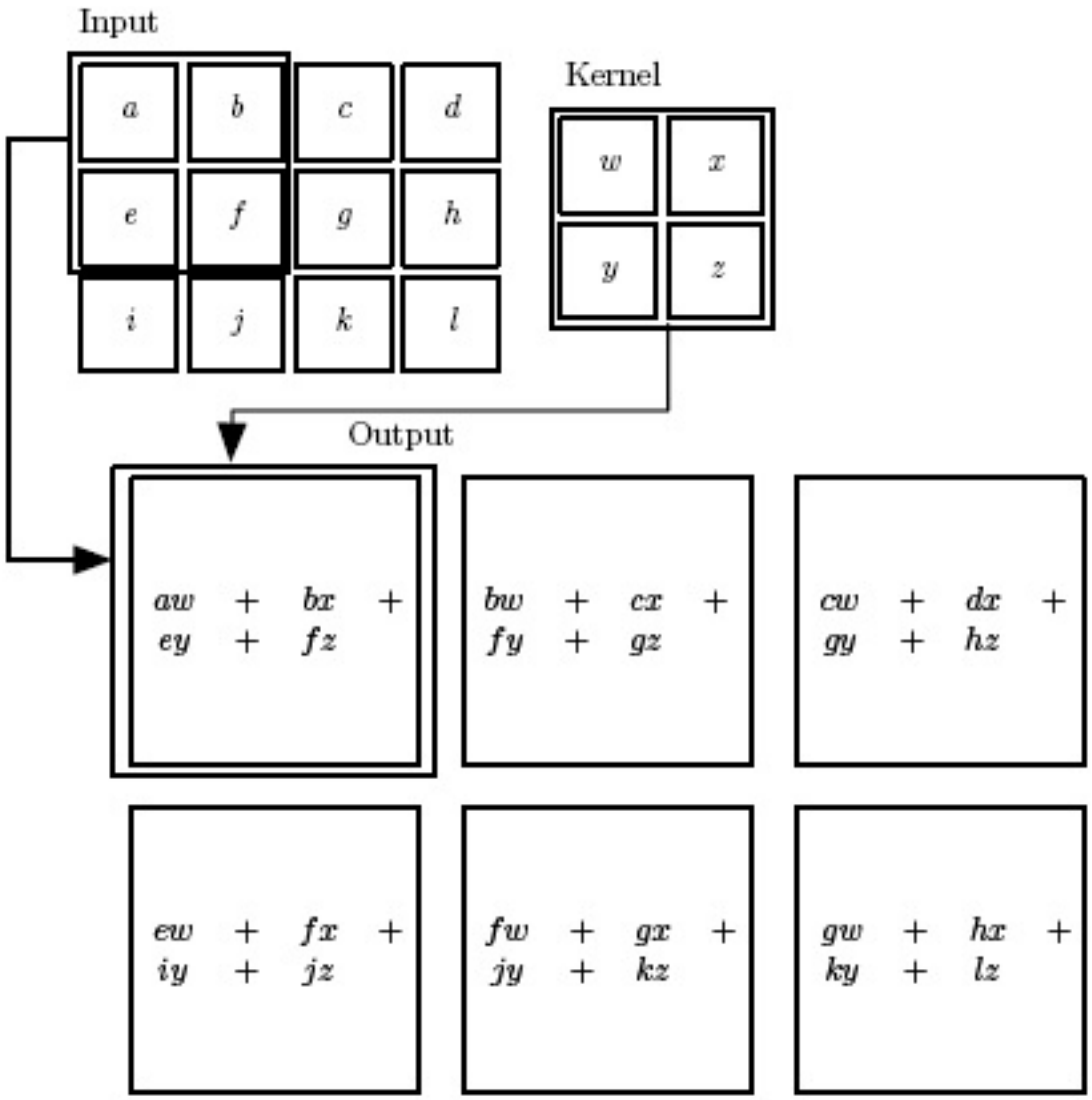


FIGURE 3.32 The mechanics of spatial filtering. The magnified drawing shows a 3×3 mask and the image section directly under it; the image section is shown displaced out from under the mask for ease of readability.

Image Processing Basics: 2D Convolution...



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)

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

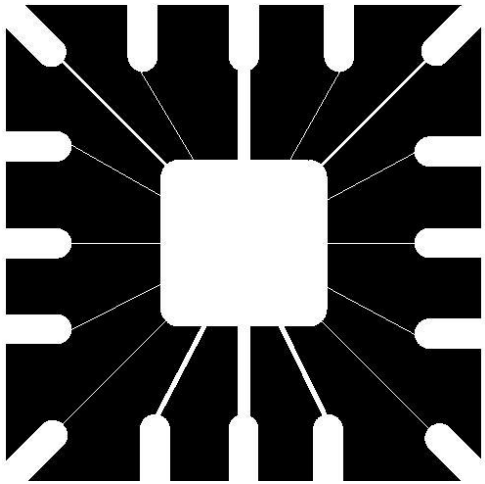
First derivative

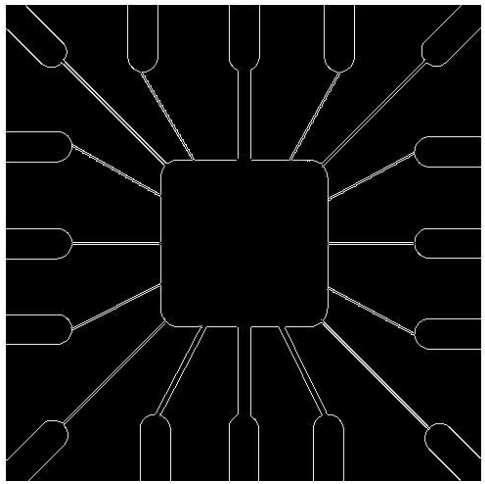
$$\frac{\partial f}{\partial x} \leftrightarrow f(x+1) - f(x)$$

Edge Detector

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

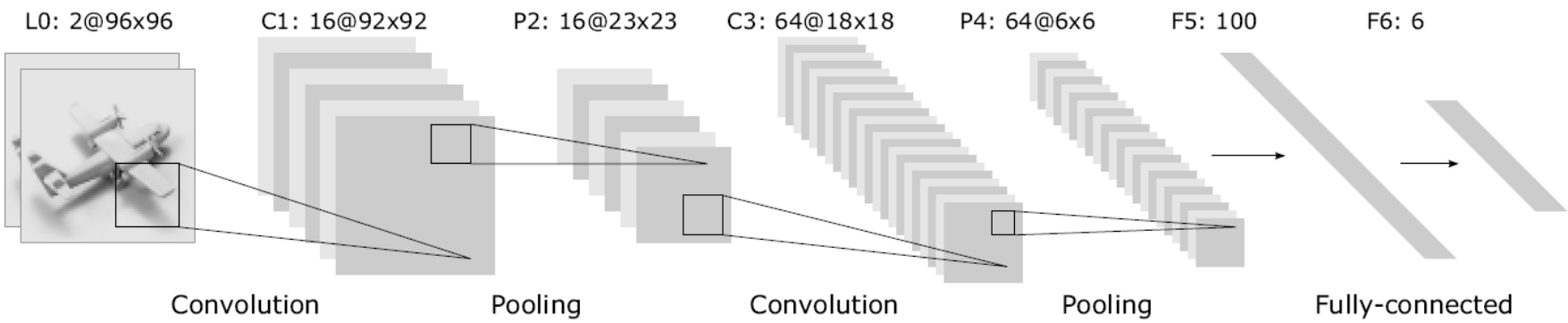
-1	-1	-1
0	0	0
1	1	1



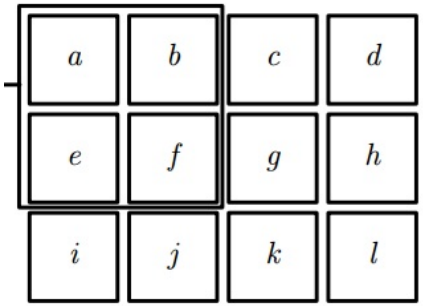


CNN

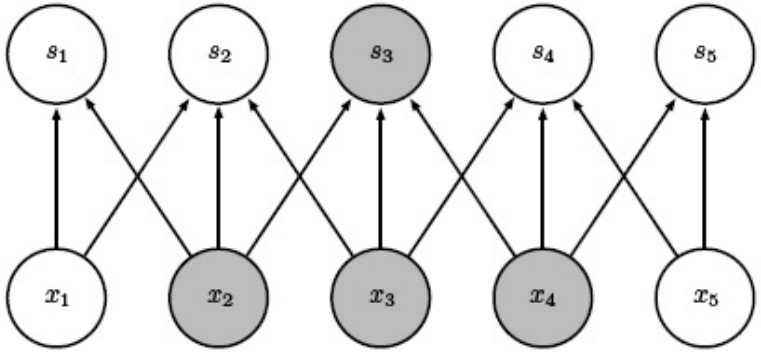
Convolutional neural network



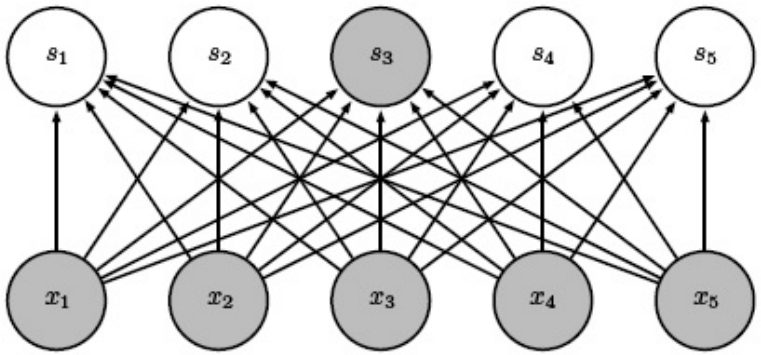
Sparse Connectivity



Sparse connections due to small convolution kernel

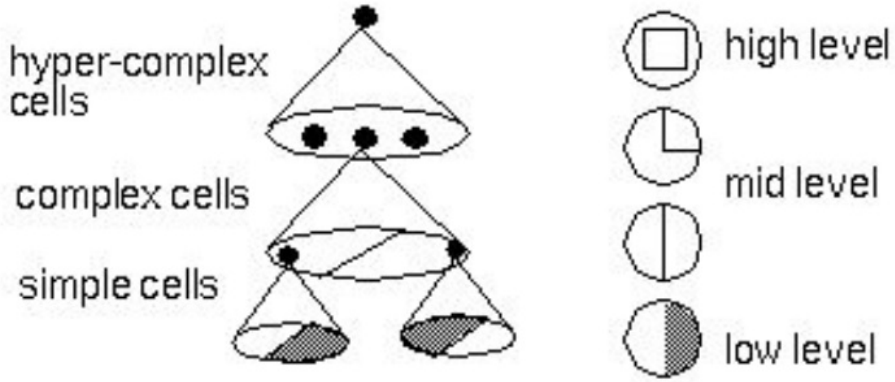


Dense connections

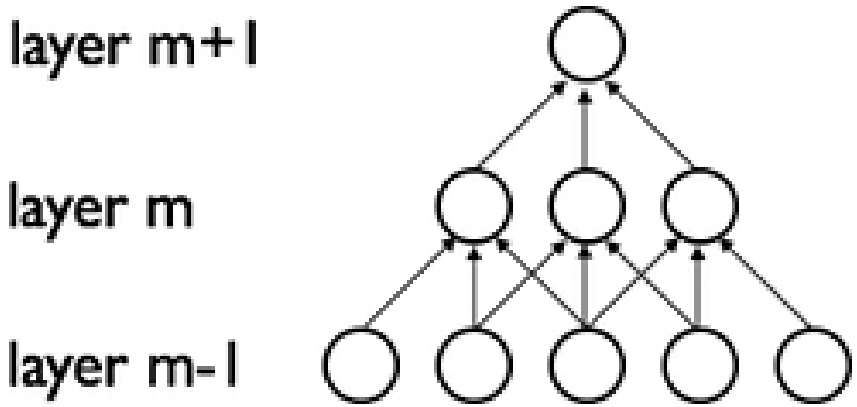


- receptive field

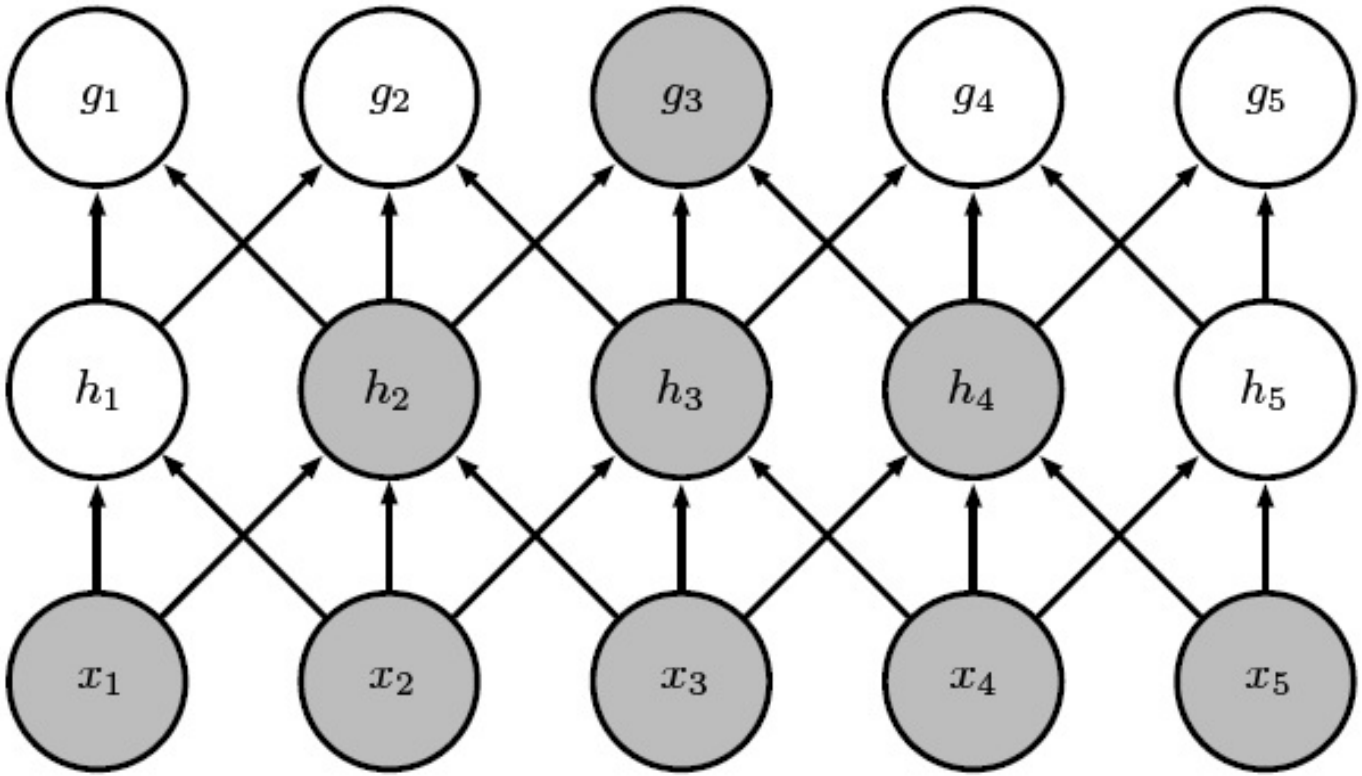
Feature Hierarchy



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

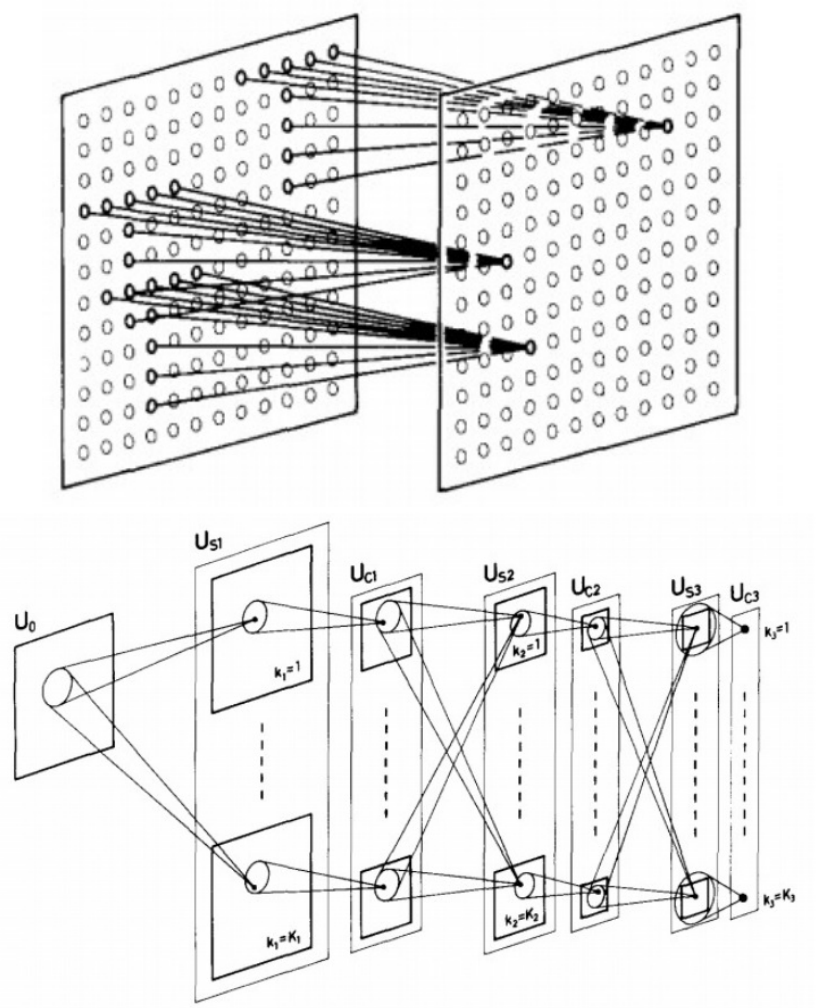


Growing Receptive Field



Feature Hierarchy...

- another early model: Neocognitron [Fukushima 1980]



Shared Weights

- each local receptive field is replicated across the entire image

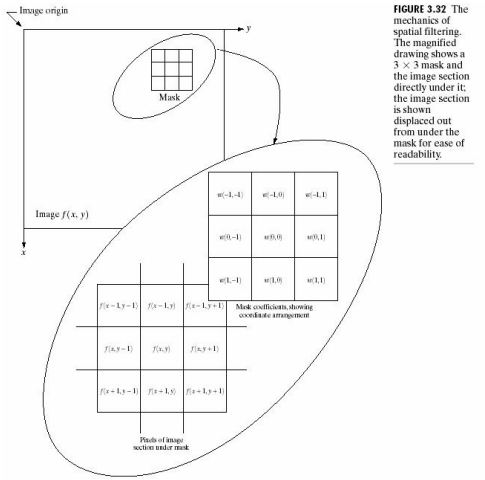
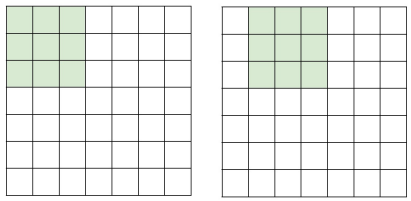
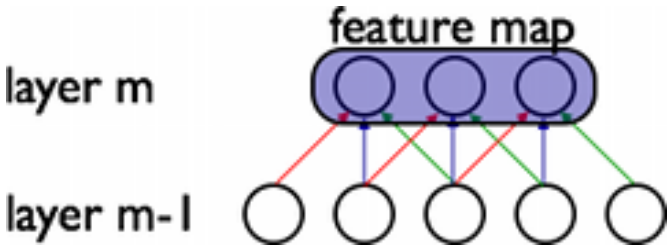


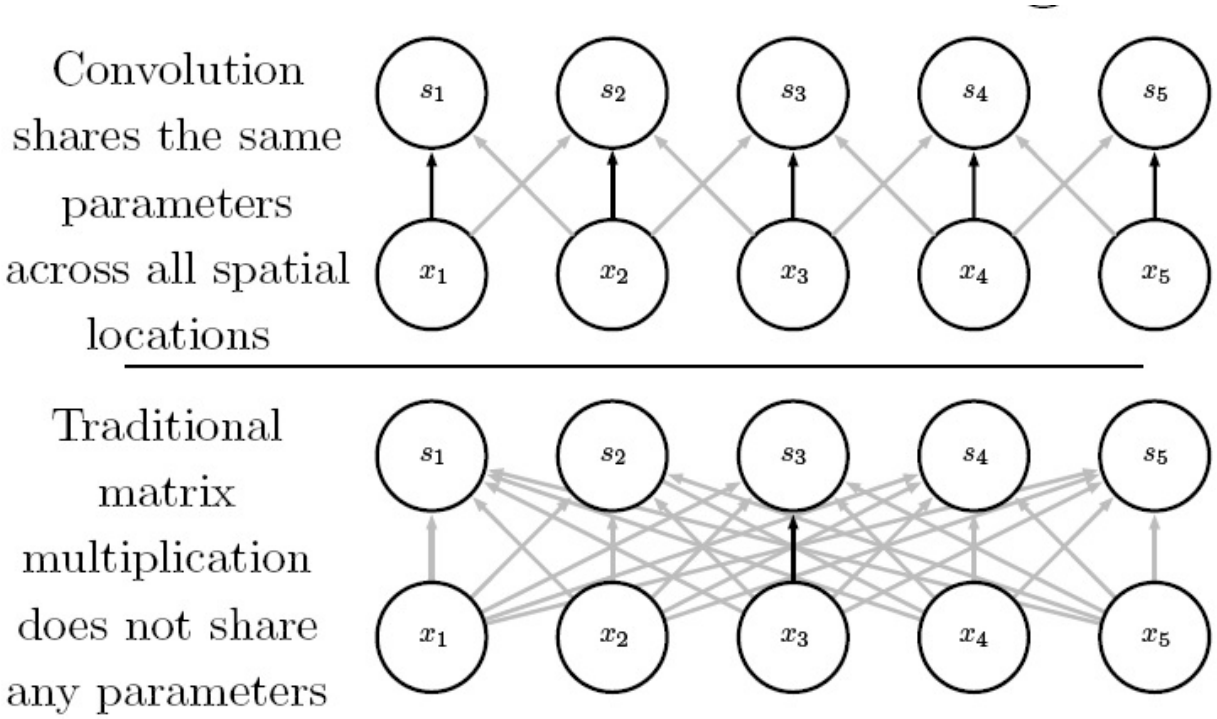
FIGURE 3.32 The mechanics of spatial filtering. The magnified drawing shows a 3×3 mask and the image section directly under it; the image section is shown displaced out from under the mask for ease of readability.



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



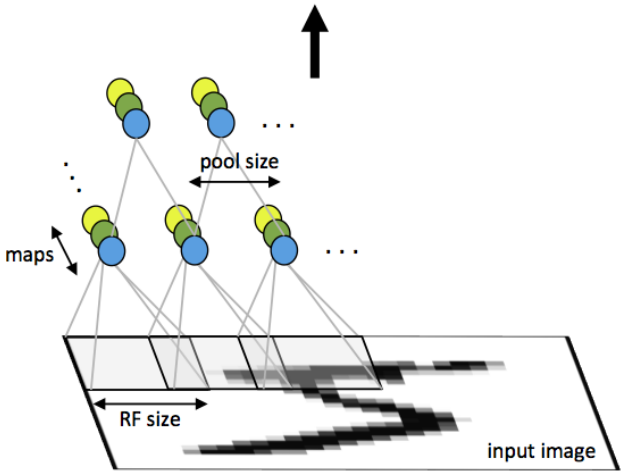
Parameter Sharing



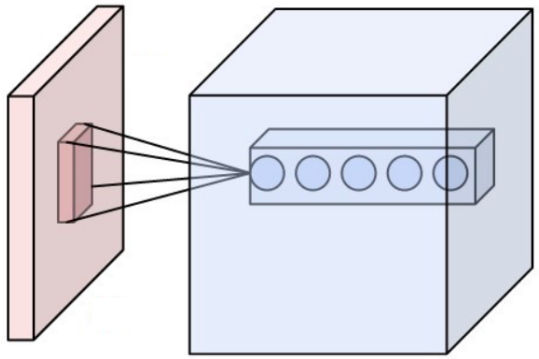
- allows for features to be detected regardless of their position in the image
  - robustness to shifts of the input
- greatly reduces the number of free parameters to learn

Convolutional Layer

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



- stack the activation maps for all filters along the depth dimension



Efficiency of Convolution

Input size: 320 by 280

Kernel size: 2 by 1

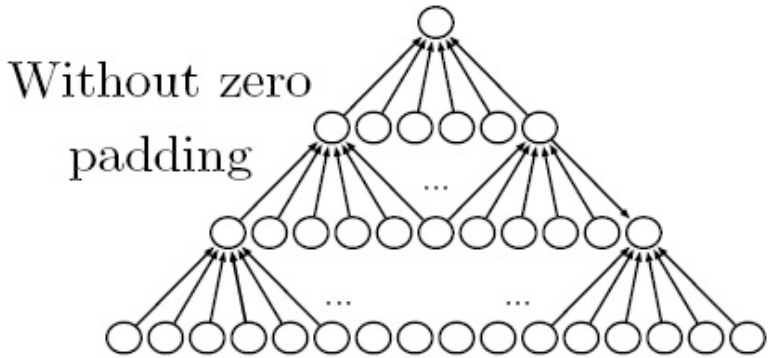
Output size: 319 by 280

	Convolution	Dense matrix	Sparse matrix
Stored floats	2	319*280*320*280 $>$ 8e9	2*319*280 = 178,640

Nonlinearity

- Convolution is a linear operation
- need nonlinearity
  - otherwise 2 convolution layers would be no more powerful than 1
- common to apply a rectified linear unit (ReLU): $y = \max(z, 0)$

Zero-Padding



- representation shrink at each layer
- limits the number of layers

Zero-padding

- adding zeros to each layer
- allows the use of an arbitrarily deep convolutional network

Pooling Layer

motivation

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

Example

the input image contains

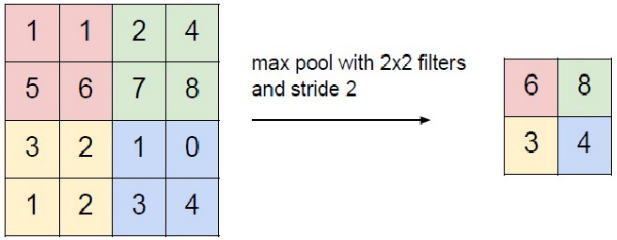
1. the endpoint of a roughly horizontal segment in the upper left area
2. a corner in the upper right area
3. the endpoint of a roughly vertical segment in the lower portion

the input image is a seven

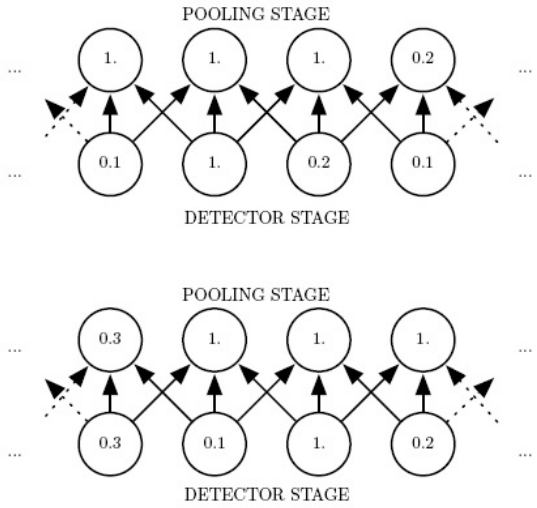
- positions are likely to vary for different instances of the character
- spatial invariance

Max-Pooling

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

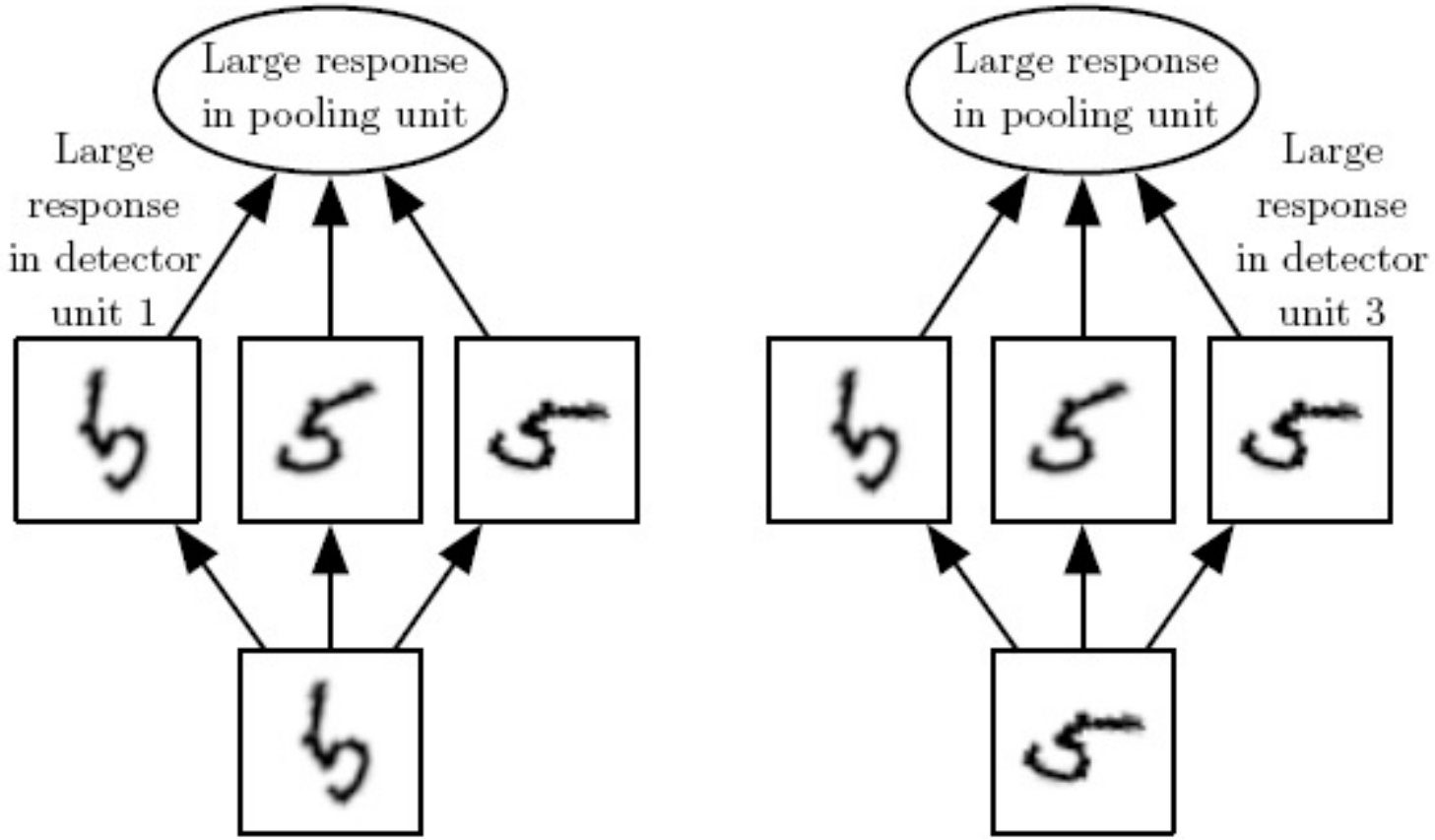


- shift the input to the right by one pixel

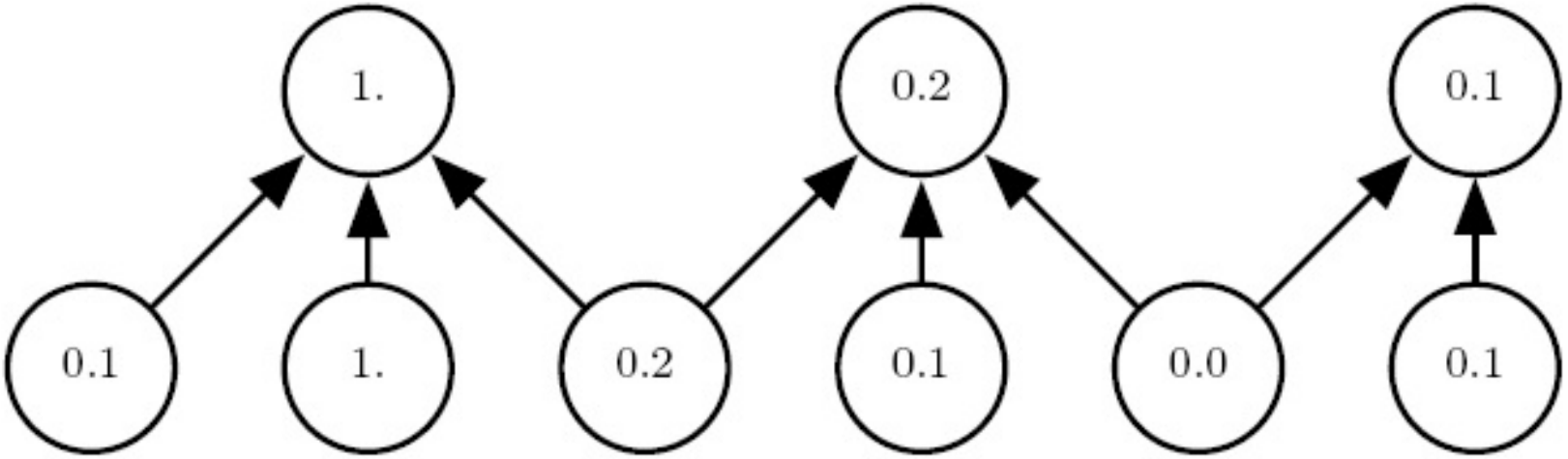


  - every value in the bottom row has changed
  - but only half of the values in the top row have changed

Example

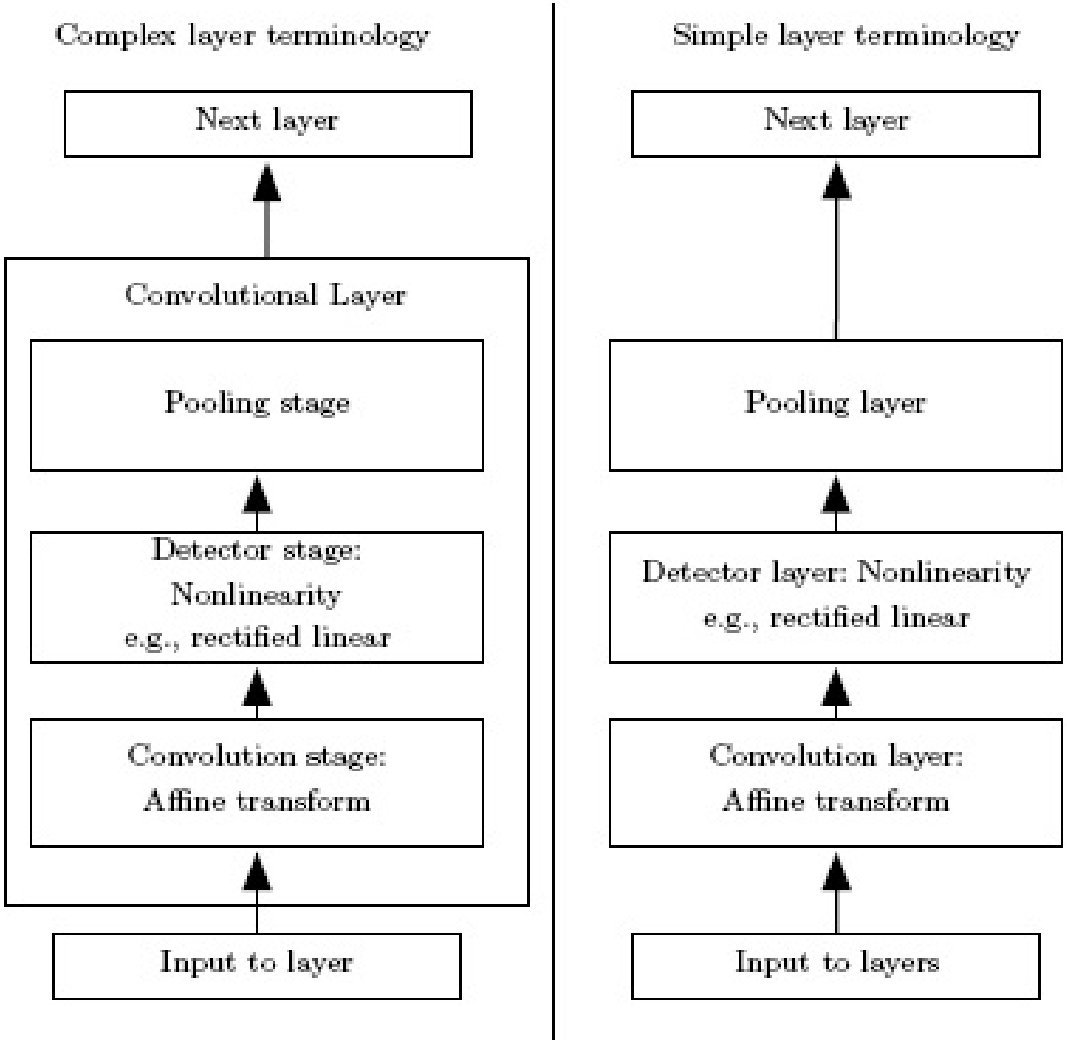

Large response
in pooling unit
Large
response
in detector
unit 1
Large response
in pooling unit
Large
response
in detector
unit 3

Pooling with Downsampling

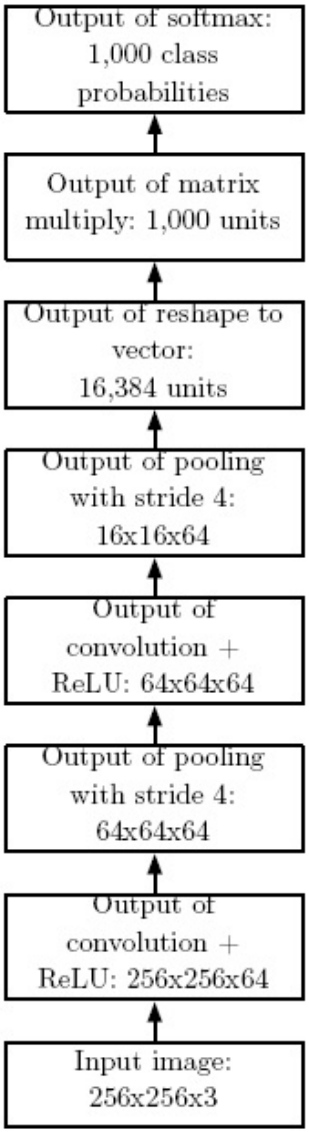


- stride of two
- reduces the representation size by a factor of two
- reduces the computational and statistical burden on the next layer

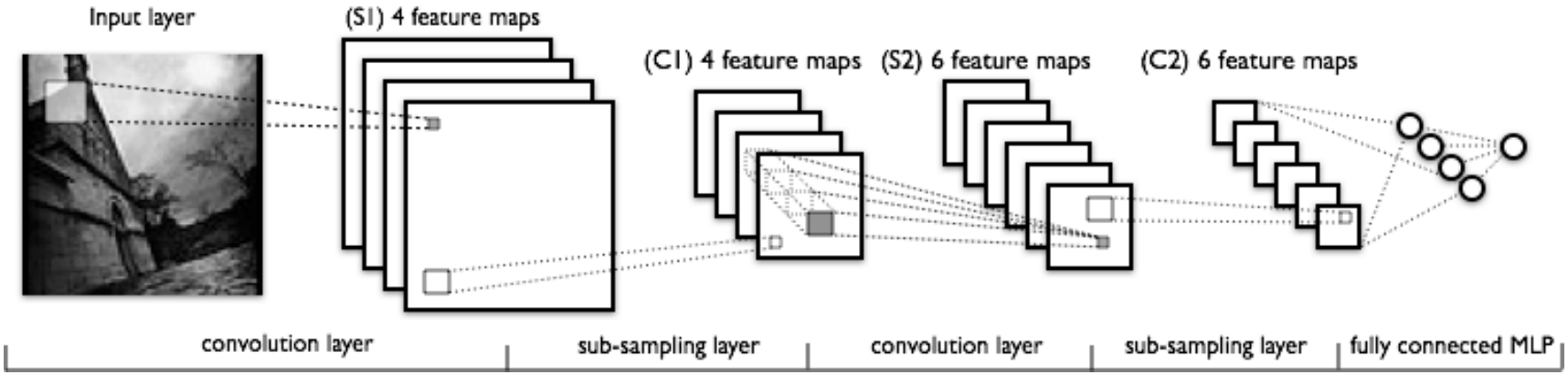
Convolutional Network Components



Example Classification Architecture



Example



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

Application: Face Recognition