Neural Networks
Part 3

Yingyu Liang
yliang@cs.wisc.edu
Computer Sciences Department
University of Wisconsin, Madison

Convolutional neural networks

- Strong empirical application performance

- Convolutional networks: neural networks that use convolution in place of general matrix multiplication in at least one of their layers

$$h = \sigma(W^T x + b)$$

for a specific kind of weight matrix W

Convolution

Convolution: discrete version

- Given array u_t and w_t, their convolution is a function s_t

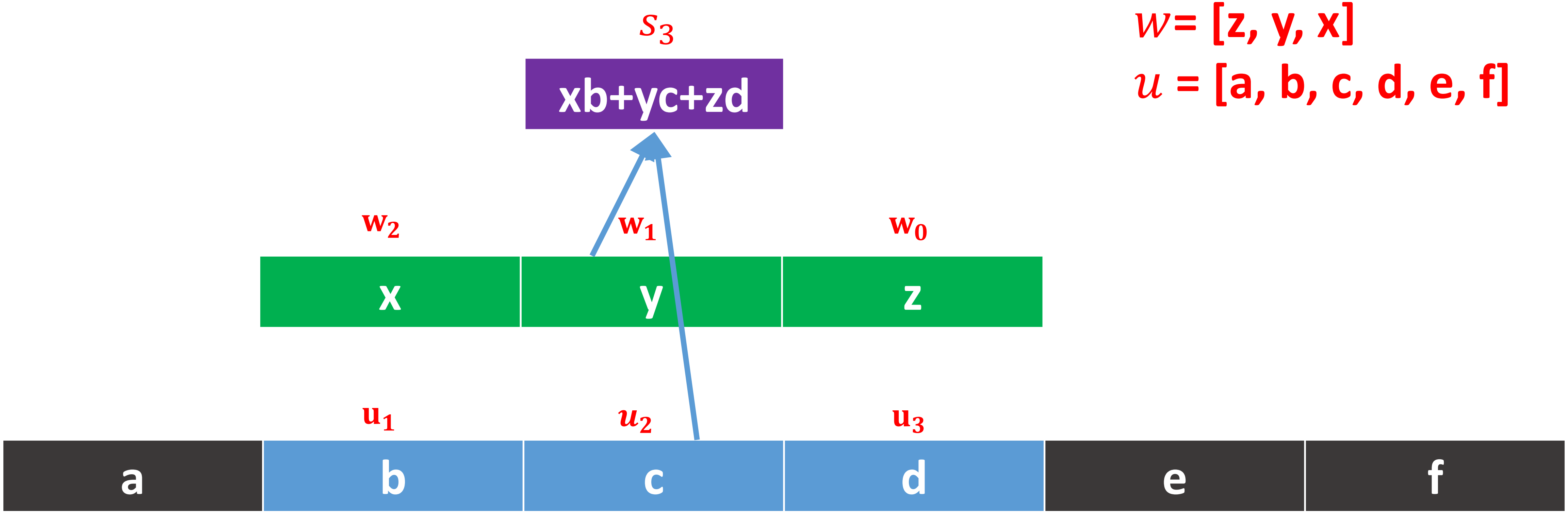
$$s_t = \sum_{a=-\infty}^{+\infty} u_a w_{t-a}$$

- Written as

$$s = (u * w) \quad \text{or} \quad s_t = (u * w)_t$$

- When u_t or w_t is not defined, assumed to be 0

Illustration 1



w= [z, y, x]

u = [a, b, c, d, e, f]

Illustration 1

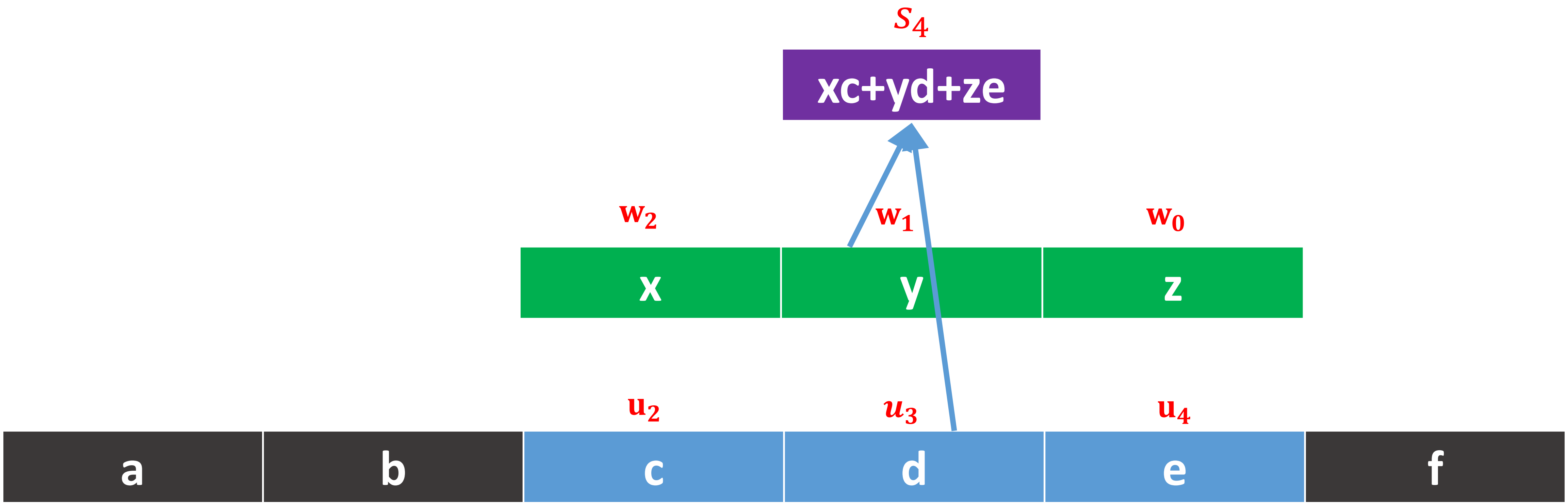


Illustration 1

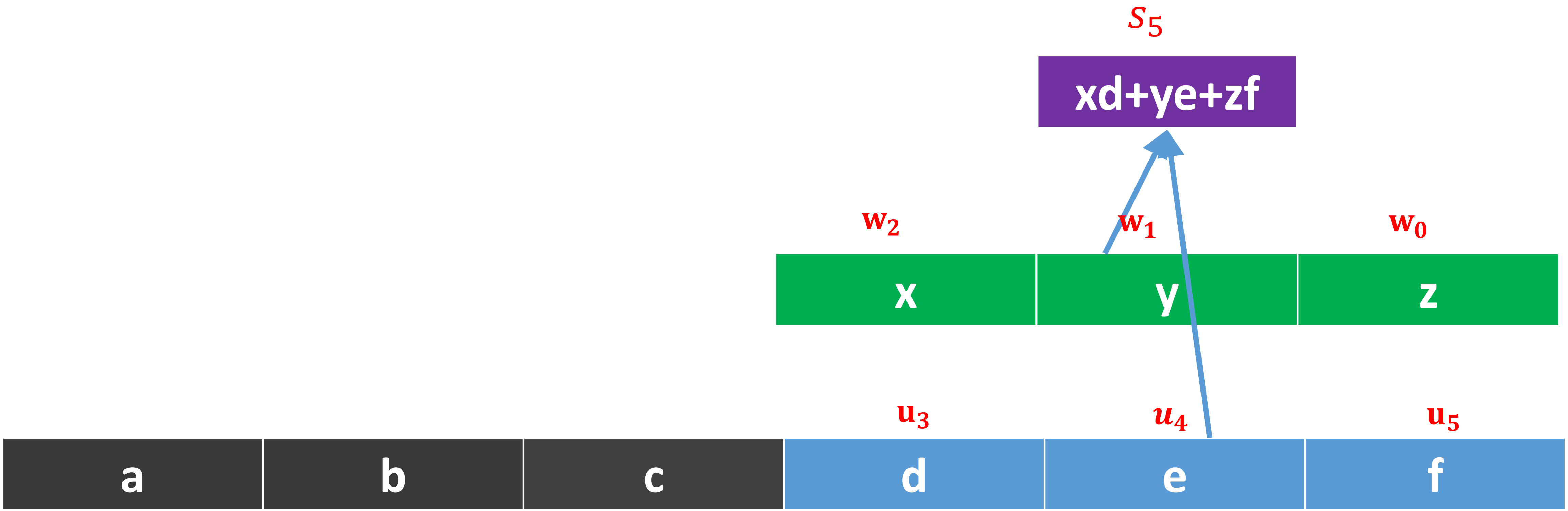


Illustration 1: boundary case

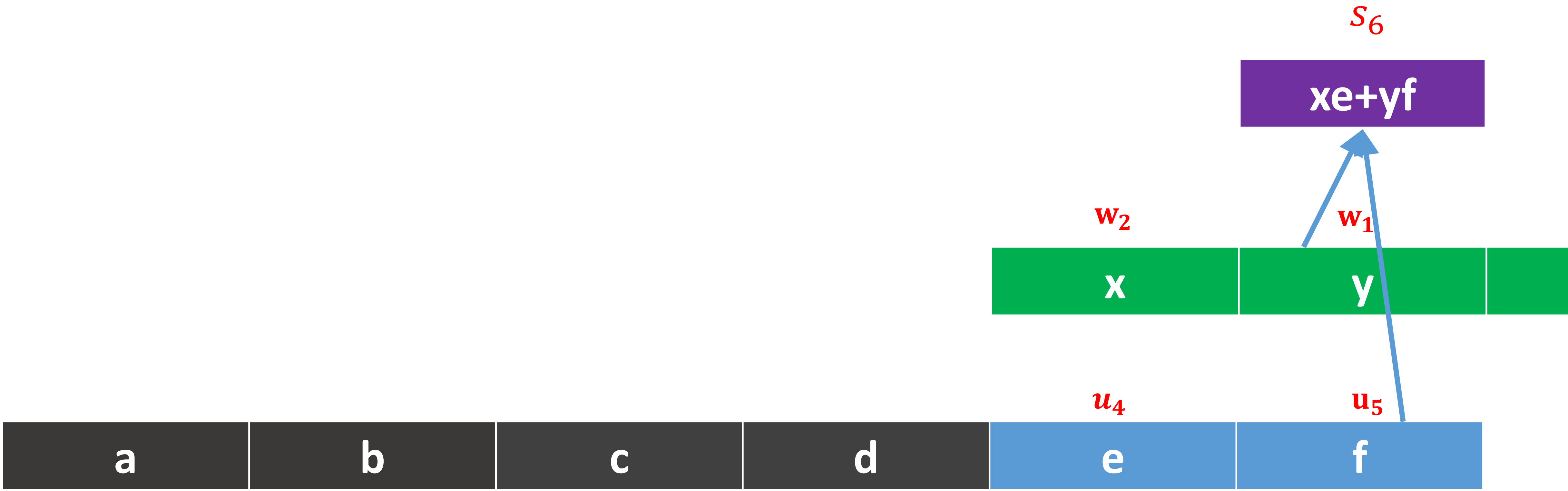


Illustration 1 as matrix multiplication

y	z				
x	y	z			
	x	y	z		
		x	y	z	
			x	y	z
				x	y

a
b
c
d
e
f

Illustration 2: two dimensional case

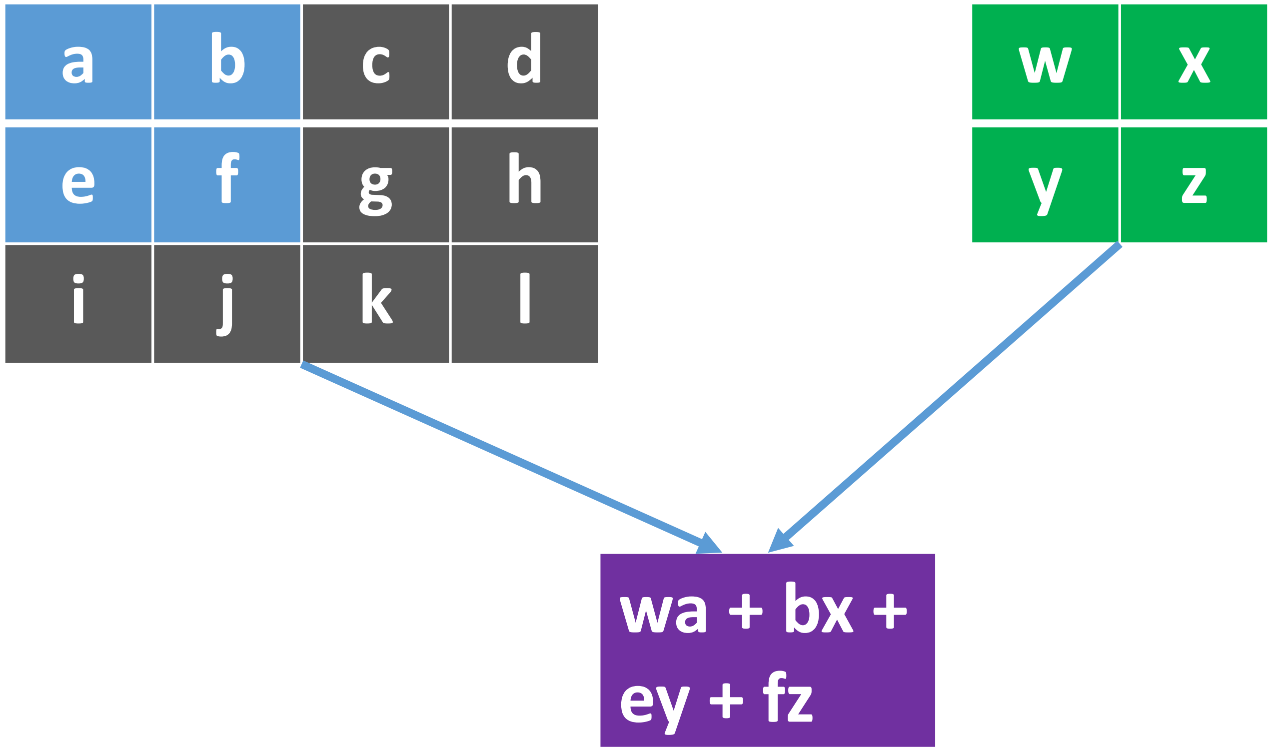


Illustration 2

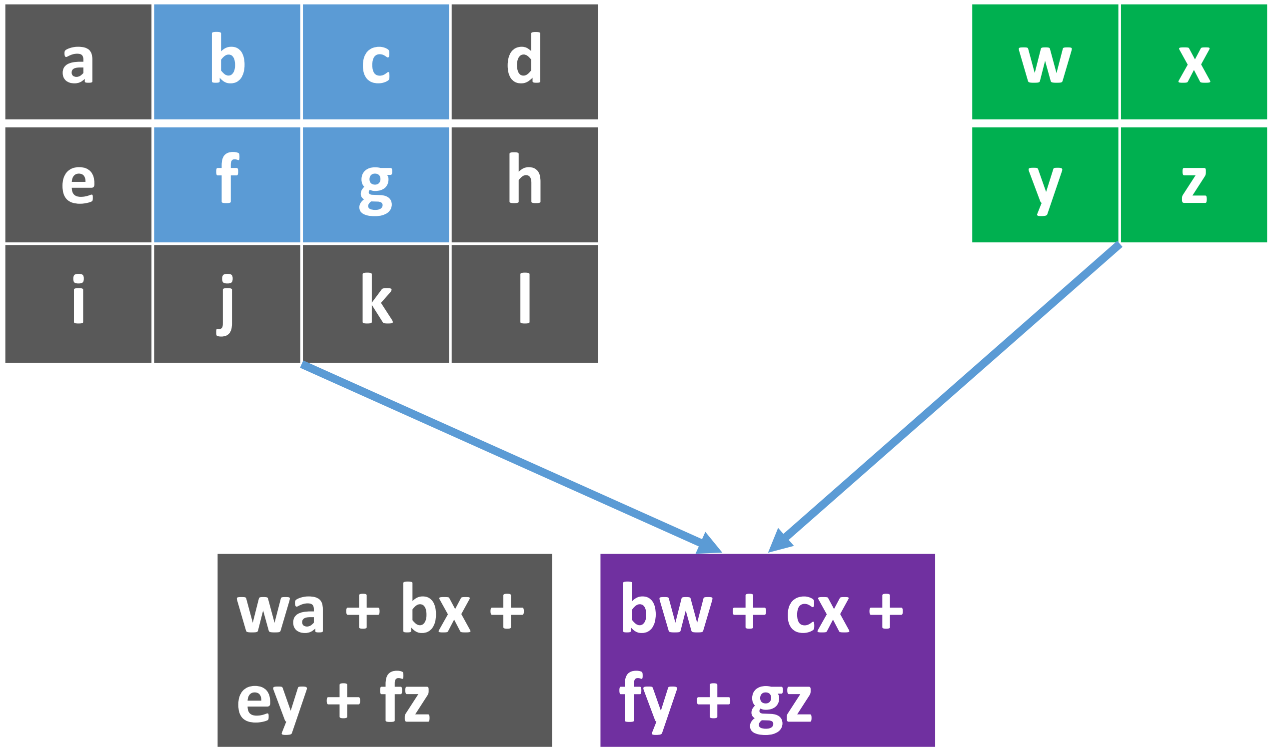
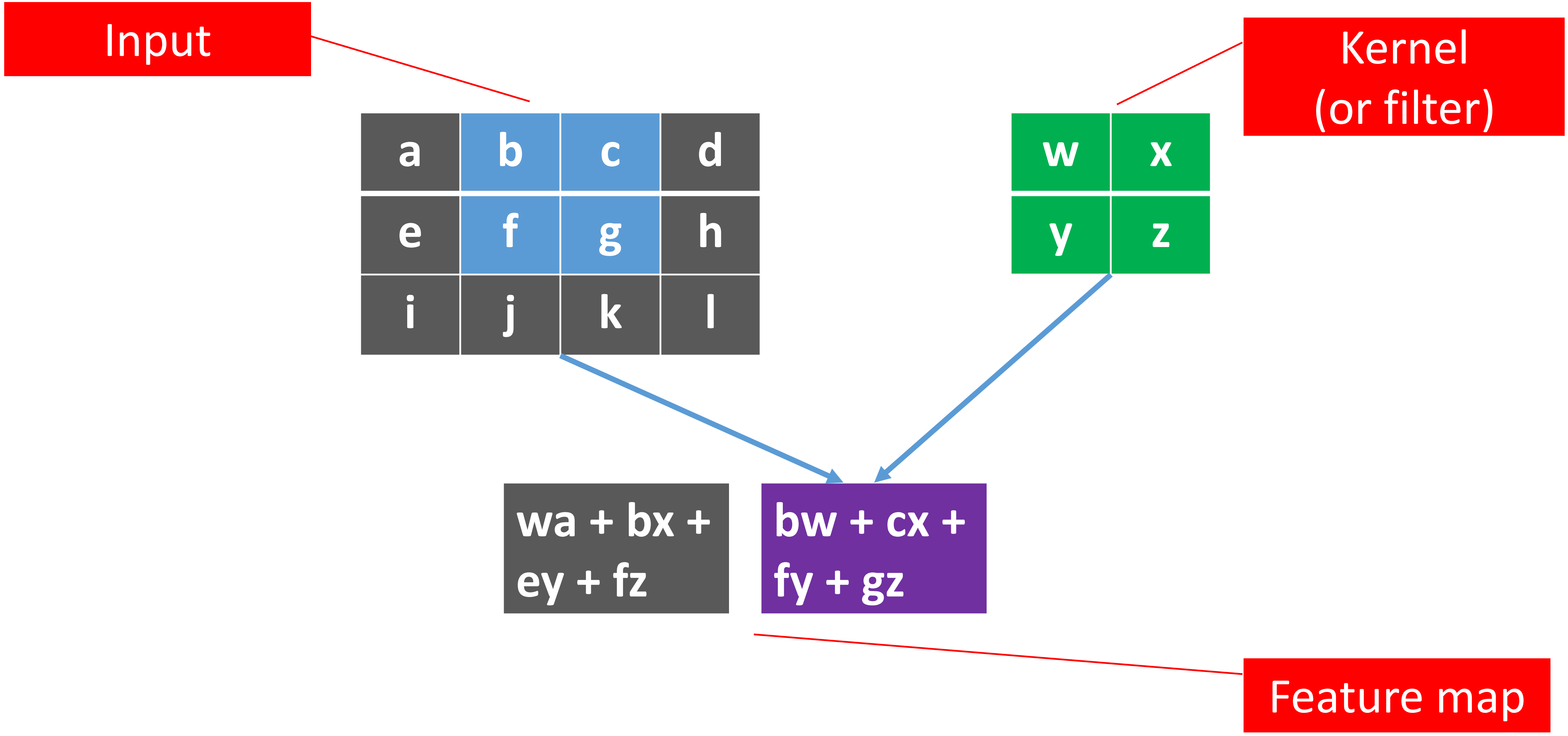
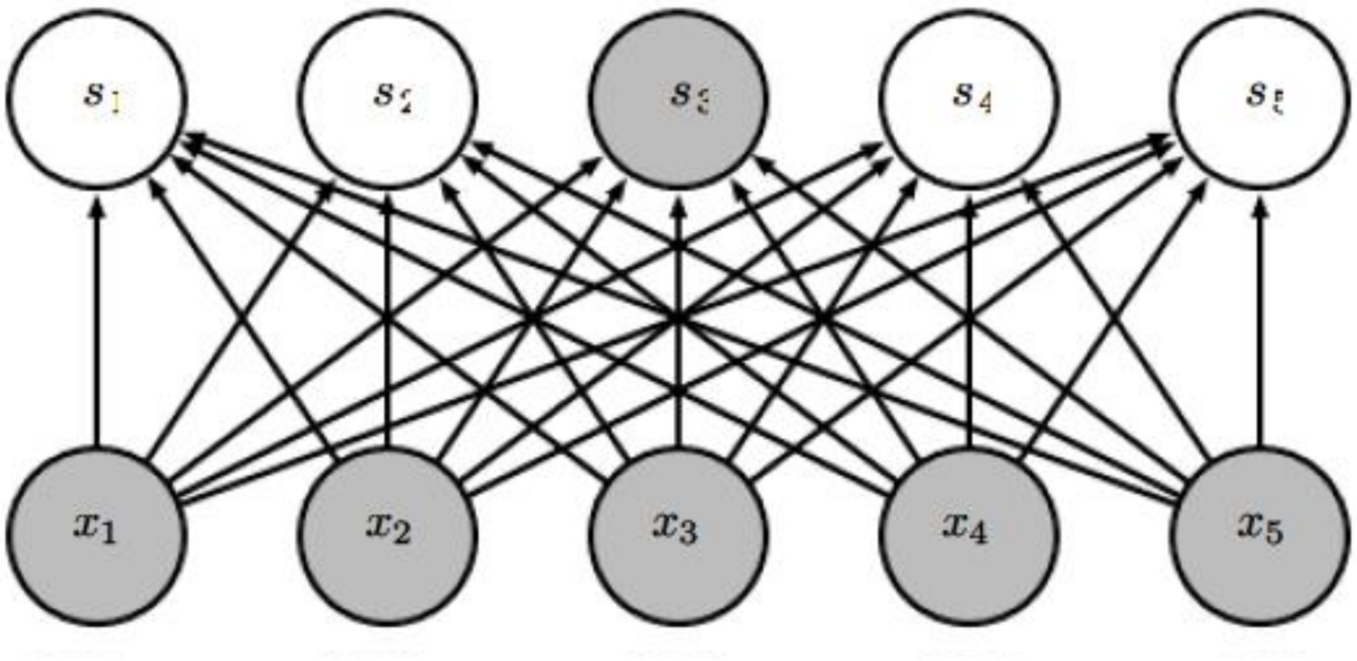


Illustration 2



Advantage: sparse interaction

Fully connected layer, $m \times n$ edges



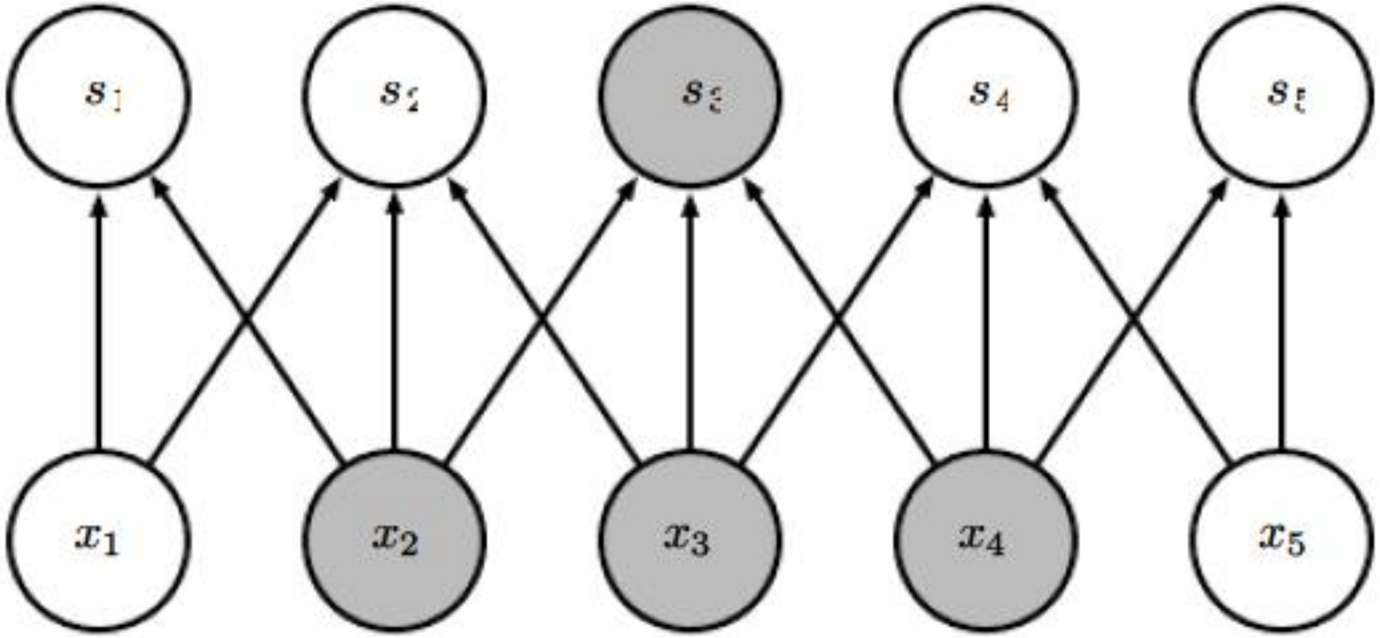
m output nodes

n input nodes

Figure from *Deep Learning,* by Goodfellow, Bengio, and Courville

Advantage: sparse interaction

Convolutional layer, $\leq m \times k$ edges



m output nodes

k kernel size

n input nodes

Figure from *Deep Learning,* by Goodfellow, Bengio, and Courville

Advantage: sparse interaction

Multiple convolutional layers: larger receptive field

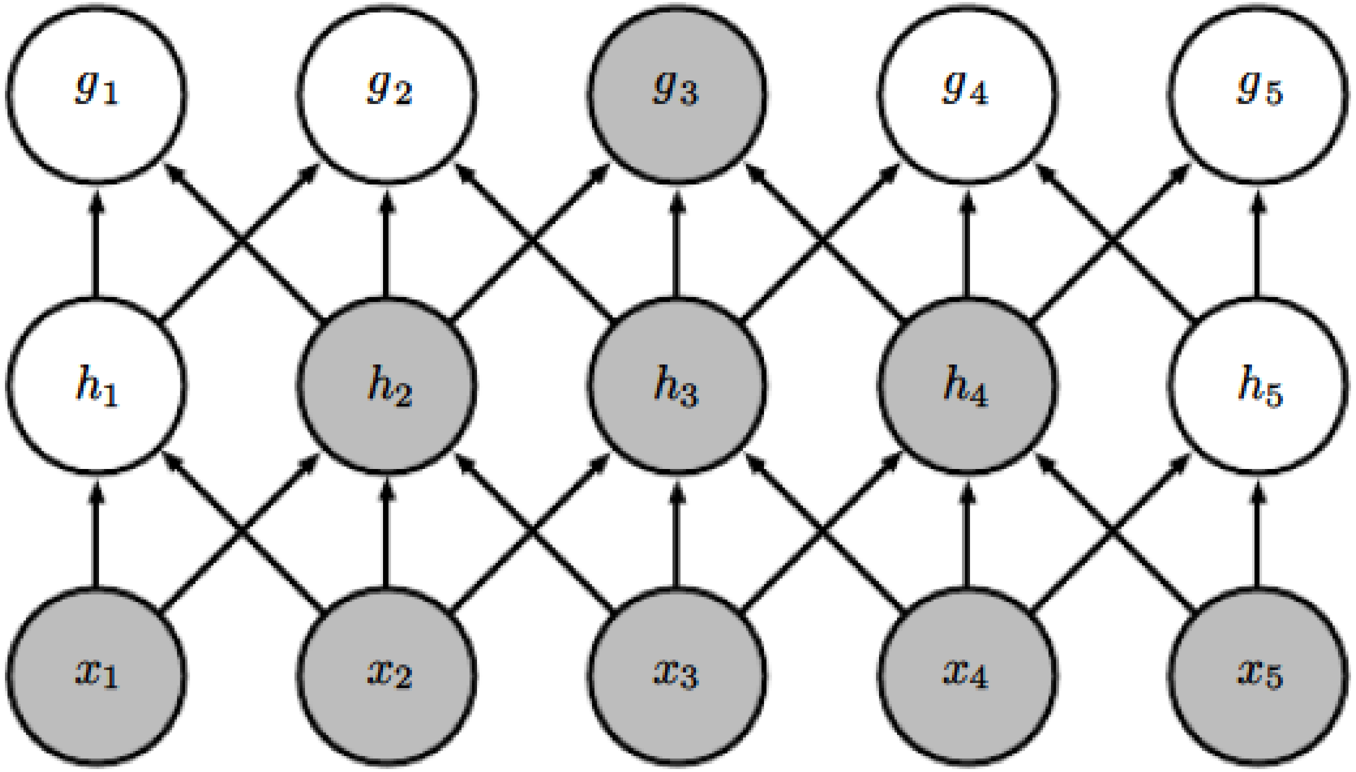
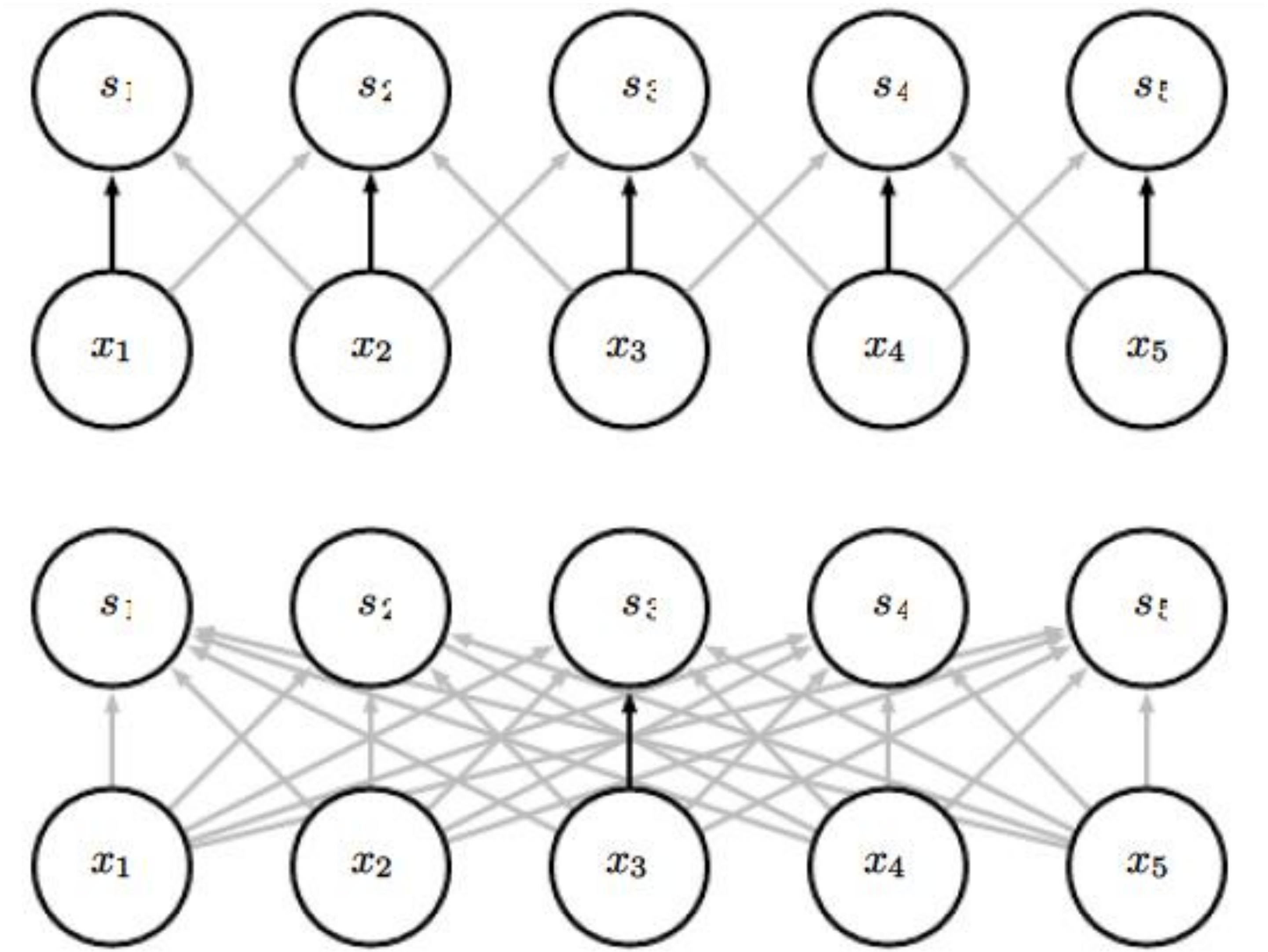


Figure from *Deep Learning,* by Goodfellow, Bengio, and Courville

Advantage: parameter sharing



The same kernel are used repeatedly. E.g., the black edge is the same weight in the kernel.

Figure from *Deep Learning,* by Goodfellow, Bengio, and Courville

Advantage: equivariant representations

- Equivariant: transforming the input = transforming the output

- Example: input is an image, transformation is shifting
- Convolution(shift(input)) = shift(Convolution(input))

- Useful when care only about the existence of a pattern, rather than the location

Pooling

- Summarizing the input (i.e., output the max of the input)

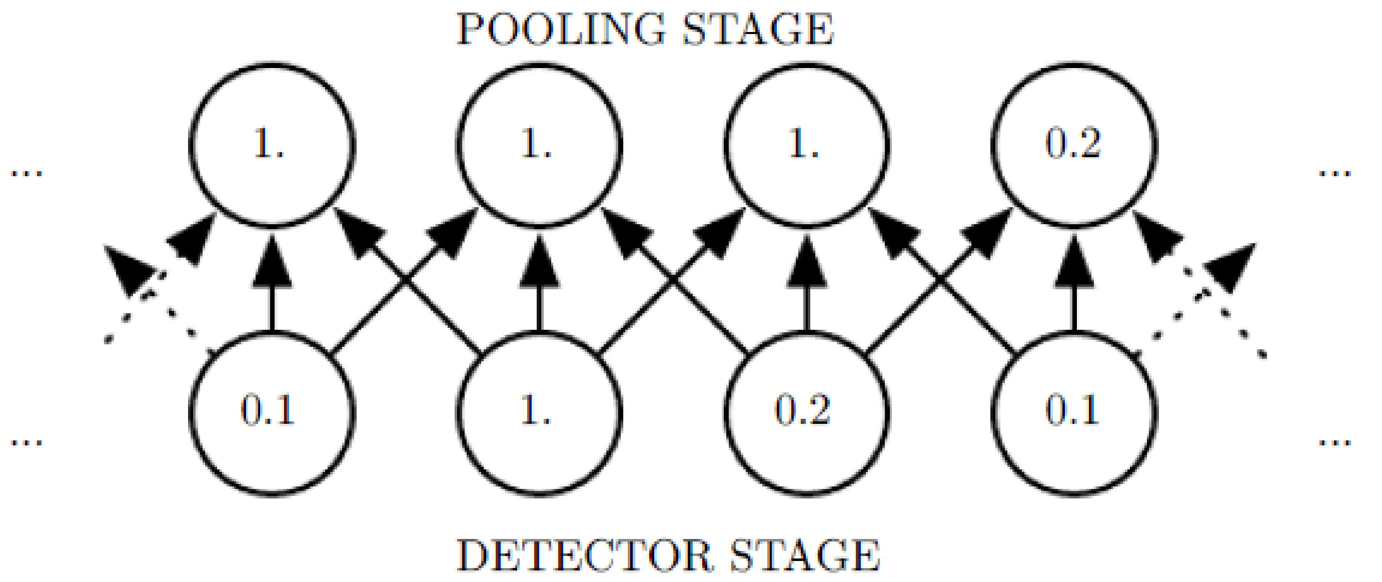


Figure from *Deep Learning,* by Goodfellow, Bengio, and Courville

Advantage

Induce invariance

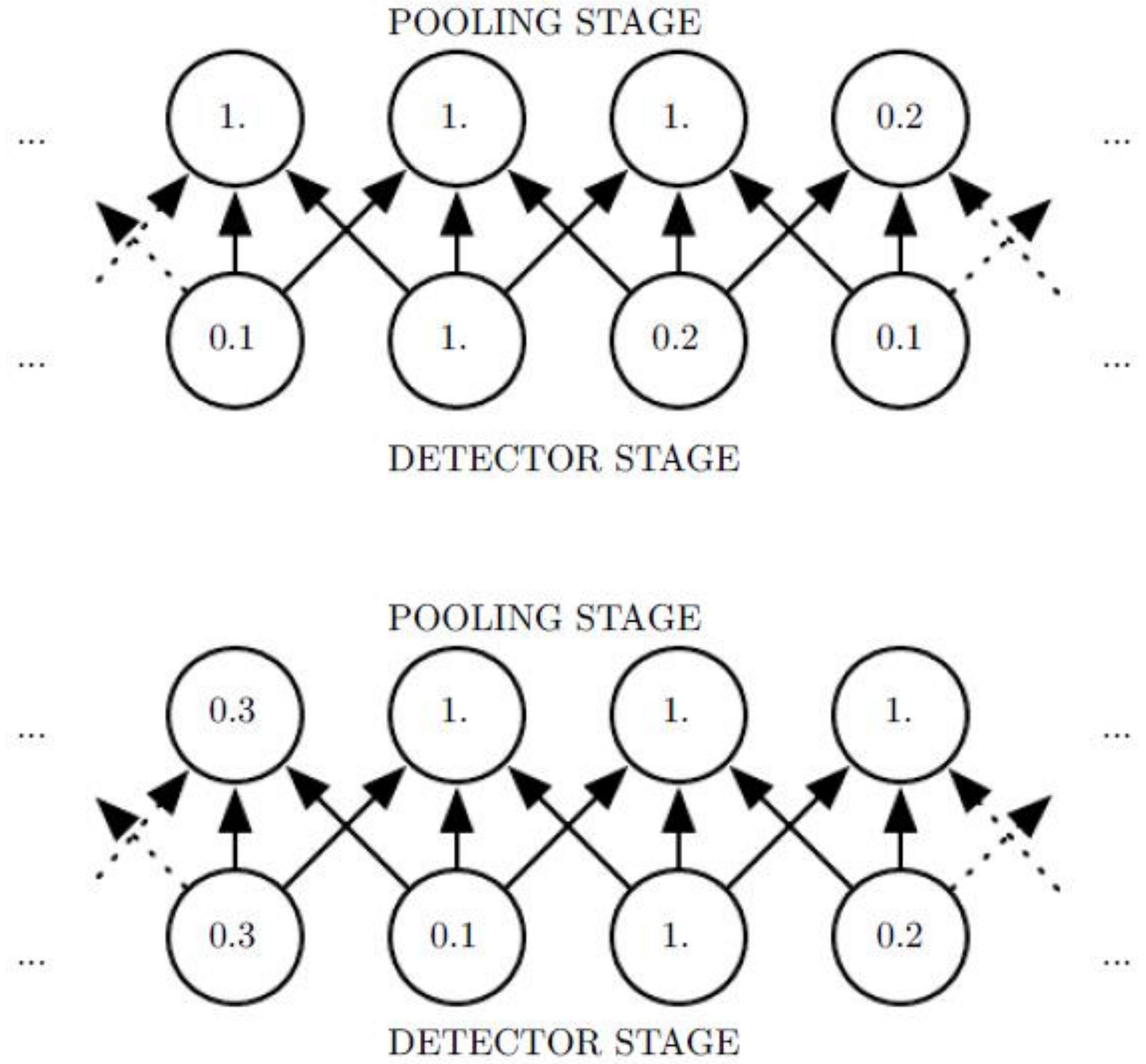


Figure from *Deep Learning,* by Goodfellow, Bengio, and Courville

Motivation from neuroscience

- David Hubel and Torsten Wiesel studied early visual system in human brain (V1 or primary visual cortex), and won Nobel prize for this

- V1 properties
  - 2D spatial arrangement
  - Simple cells: inspire convolution layers
  - Complex cells: inspire pooling layers

Variants of convolution and pooling

Variants of convolutional layers

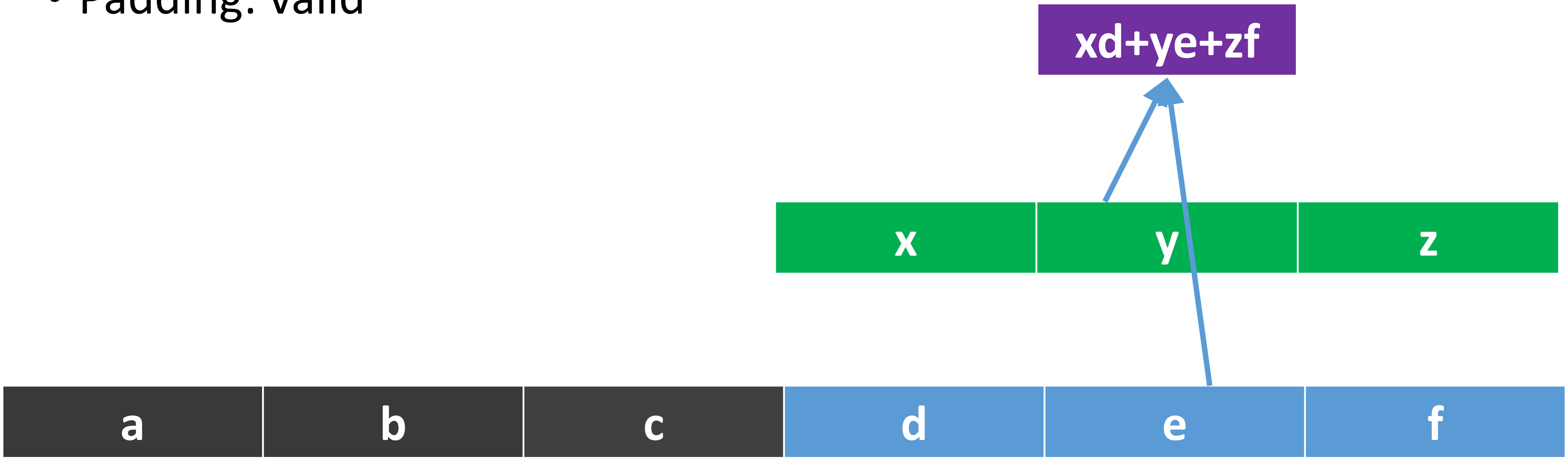
- Multiple dimensional convolution

- Input and kernel can be 3D
  - E.g., images have (width, height, RBG channels)
- Multiple kernels lead to multiple feature maps (also called channels)

- Mini-batch of images have 4D: (image_id, width, height, RBG channels)

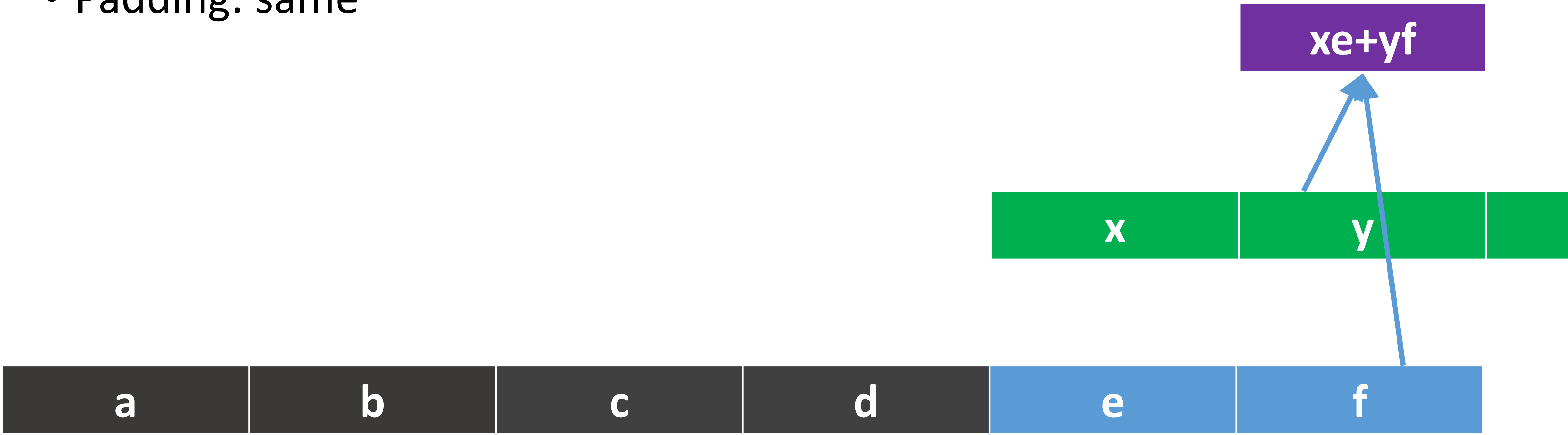
Variants of convolutional layers

- Padding: valid



Variants of convolutional layers

- Padding: same



Variants of convolutional layers

- Stride

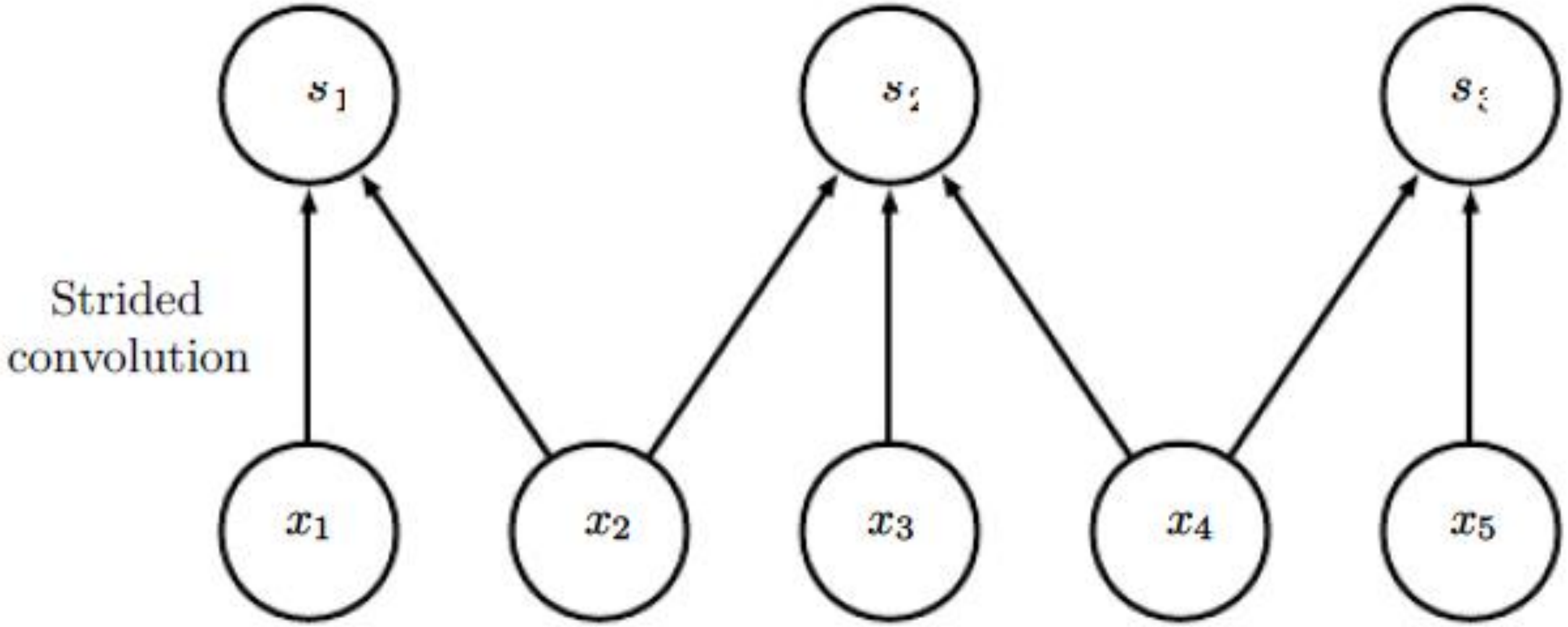


Figure from *Deep Learning,* by Goodfellow, Bengio, and Courville

Variants of pooling

- Stride and padding

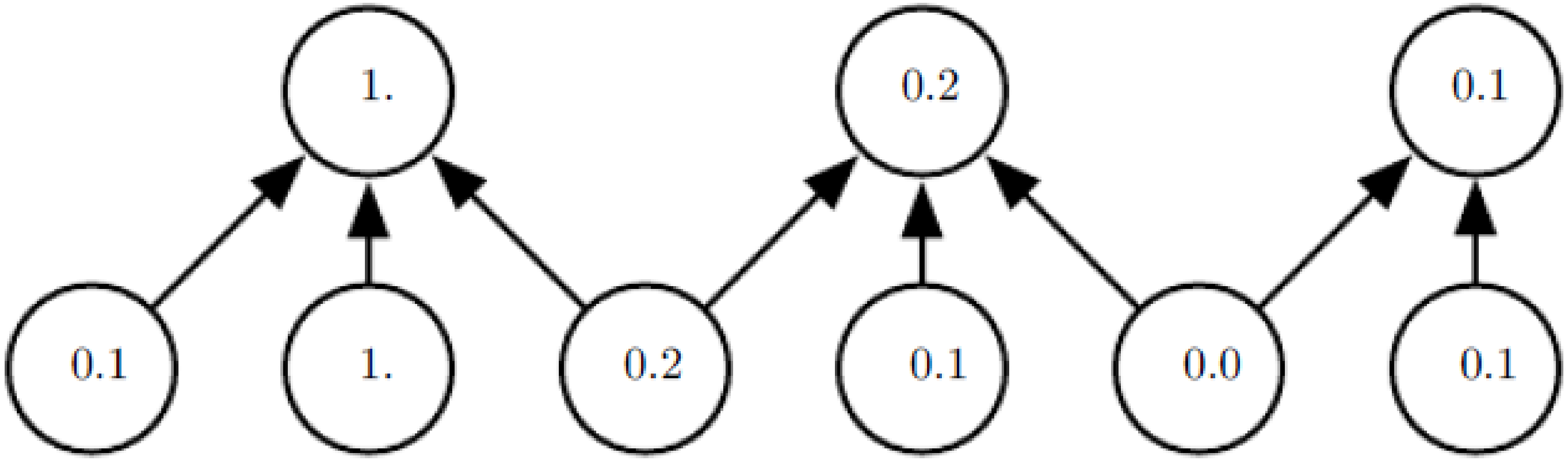


Figure from *Deep Learning,* by Goodfellow, Bengio, and Courville

Variants of pooling

- Max pooling $y = \max\{x_1, x_2, \ldots, x_k\}$
- Average pooling $y = \mathrm{mean}\{x_1, x_2, \ldots, x_k\}$

- Others like max-out

Case study: LeNet-5

LeNet-5

- Proposed in *“Gradient-based learning applied to document recognition”* , *by Yann LeCun, Leon Bottou, Yoshua Bengio and Patrick Haffner, in Proceedings of the IEEE, 1998*

LeNet-5

- Proposed in *"Gradient-based learning applied to document recognition"* , *by Yann LeCun, Leon Bottou, Yoshua Bengio and Patrick Haffner, in Proceedings of the IEEE, 1998*

- Apply convolution on 2D images (MNIST) and use backpropagation

LeNet-5

- Proposed in *"Gradient-based learning applied to document recognition"*, *by Yann LeCun, Leon Bottou, Yoshua Bengio and Patrick Haffner, in Proceedings of the IEEE, 1998*

- Apply convolution on 2D images (MNIST) and use backpropagation

- Structure: 2 convolutional layers (with pooling) + 3 fully connected layers
  - Input size: 32x32x1
  - Convolution kernel size: 5x5
  - Pooling: 2x2

LeNet-5

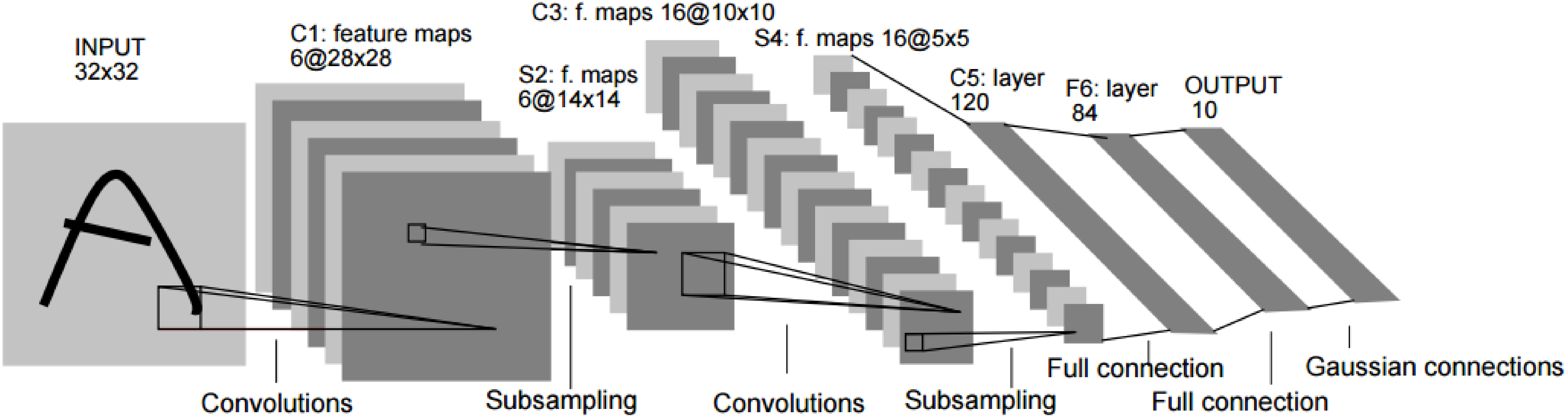


Figure from *Gradient-based learning applied to document recognition, by Y. LeCun, L. Bottou, Y. Bengio and P. Haffner*

LeNet-5

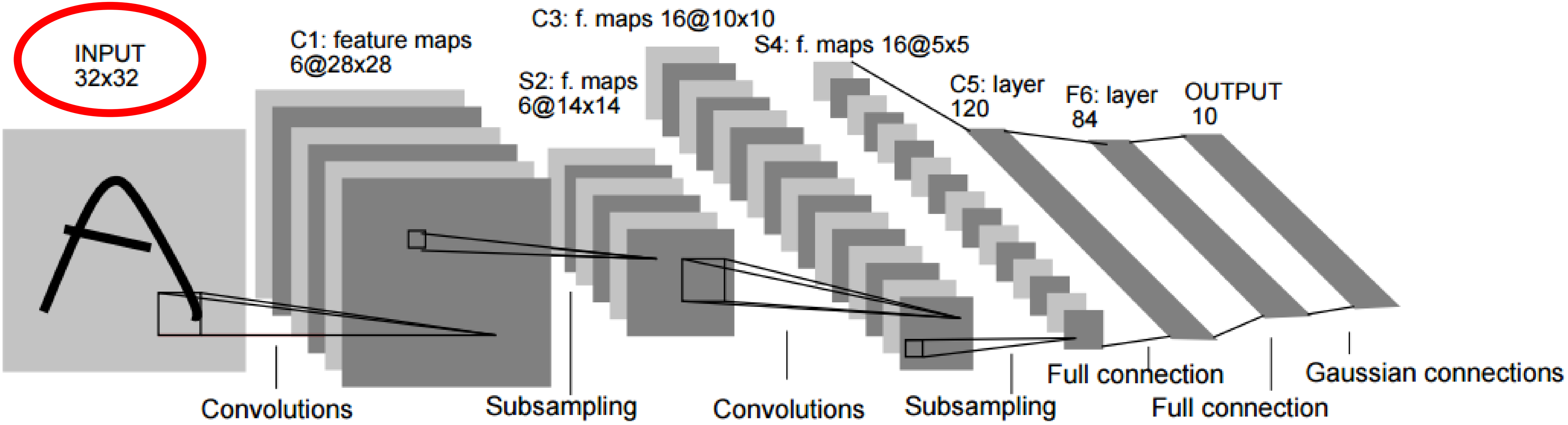


Figure from *Gradient-based learning applied to document recognition, by Y. LeCun, L. Bottou, Y. Bengio and P. Haffner*

LeNet-5

Filter: 5x5, stride: 1x1, #filters: 6

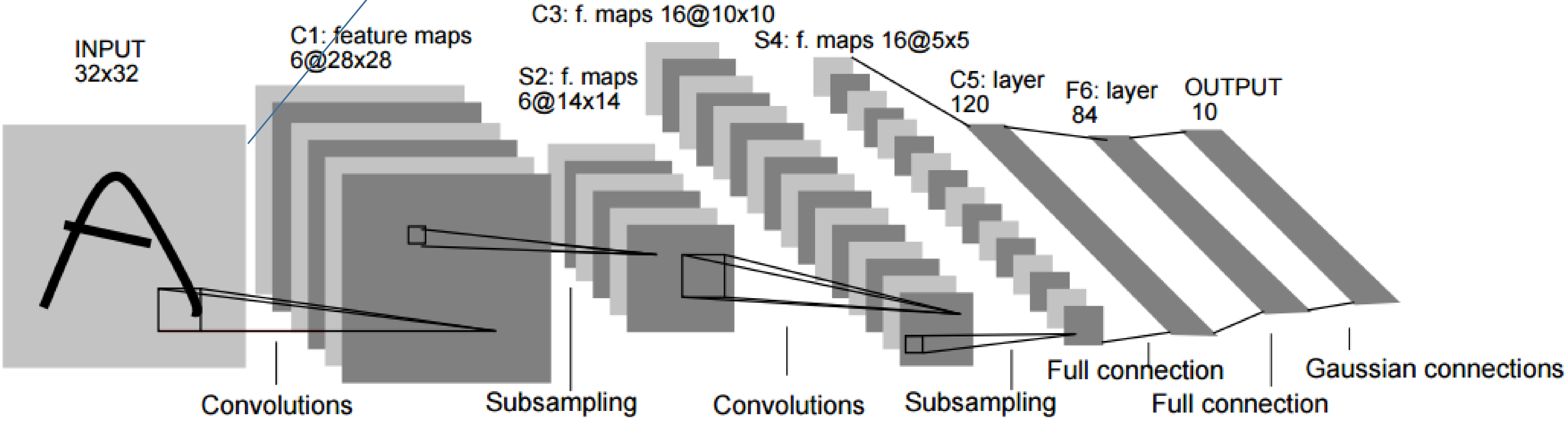


Figure from *Gradient-based learning applied to document recognition, by Y. LeCun, L. Bottou, Y. Bengio and P. Haffner*

LeNet-5

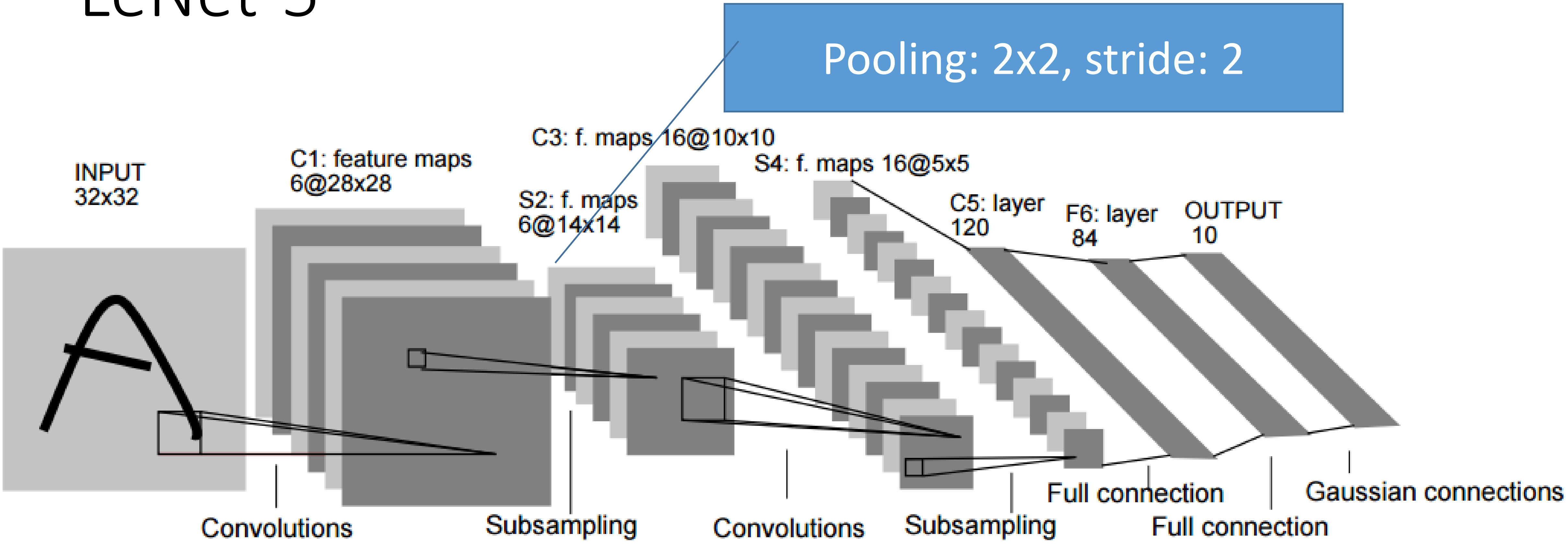


Figure from *Gradient-based learning applied to document recognition, by Y. LeCun, L. Bottou, Y. Bengio and P. Haffner*

LeNet-5

Filter: 5x5x6, stride: 1x1,
#filters: 16

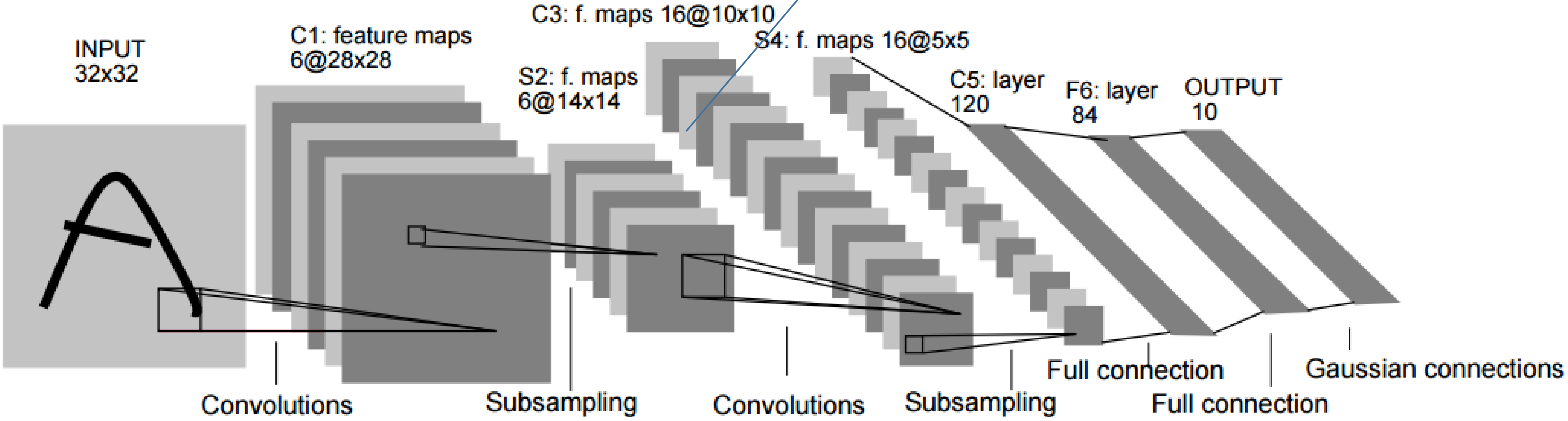


Figure from *Gradient-based learning applied to document recognition, by Y. LeCun, L. Bottou, Y. Bengio and P. Haffner*

LeNet-5

Pooling: 2x2, stride: 2

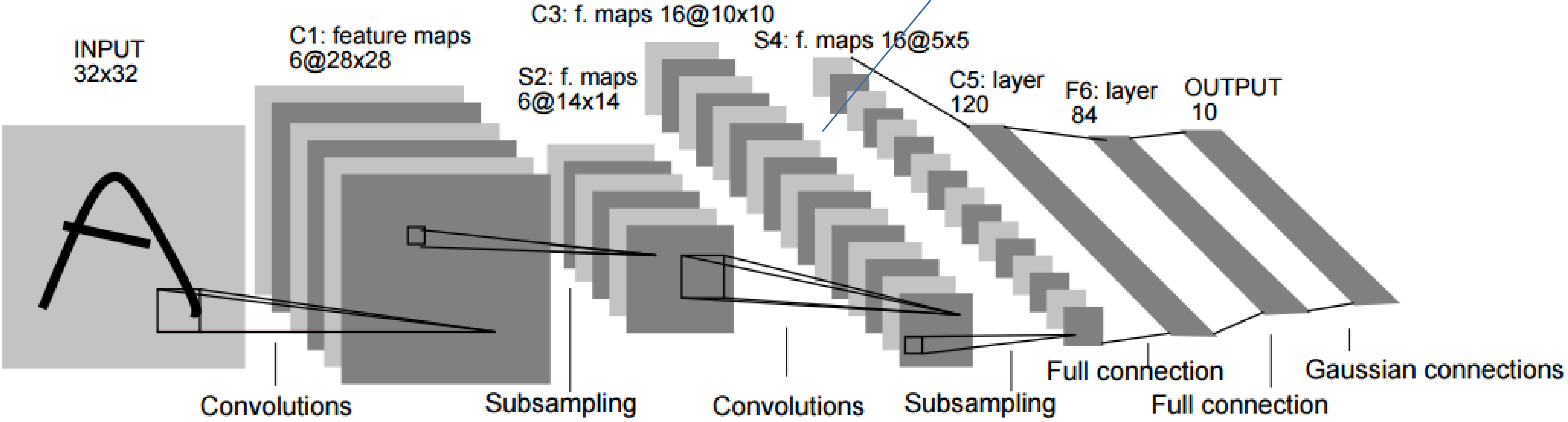


Figure from *Gradient-based learning applied to document recognition, by Y. LeCun, L. Bottou, Y. Bengio and P. Haffner*

LeNet-5

Weight matrix: 400x120

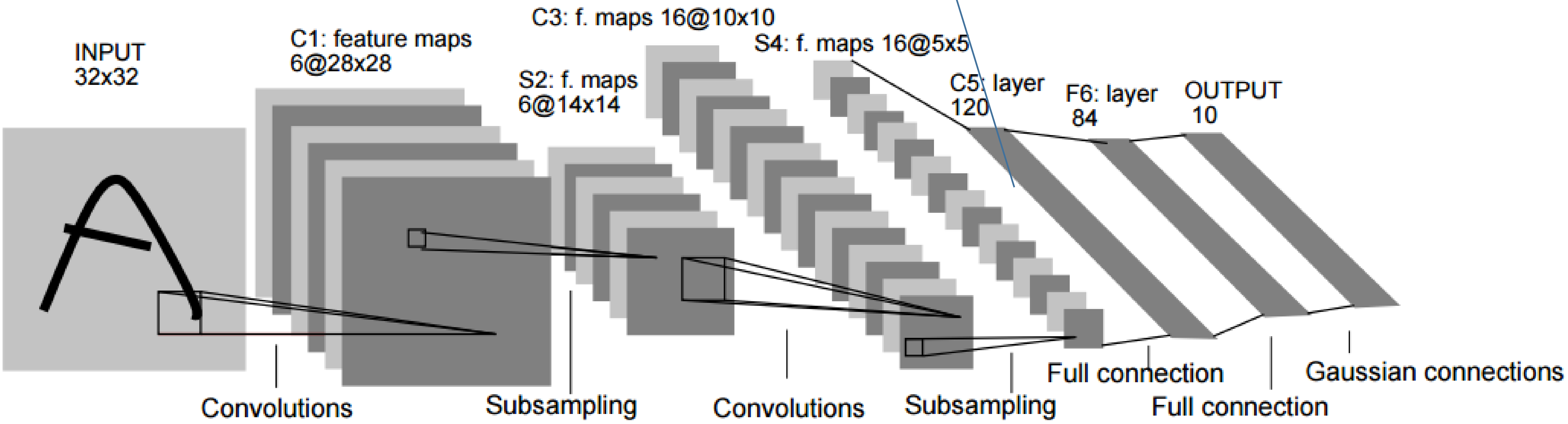


Figure from *Gradient-based learning applied to document recognition, by Y. LeCun, L. Bottou, Y. Bengio and P. Haffner*

LeNet-5

Weight matrix: 84x10

Weight matrix: 120x84

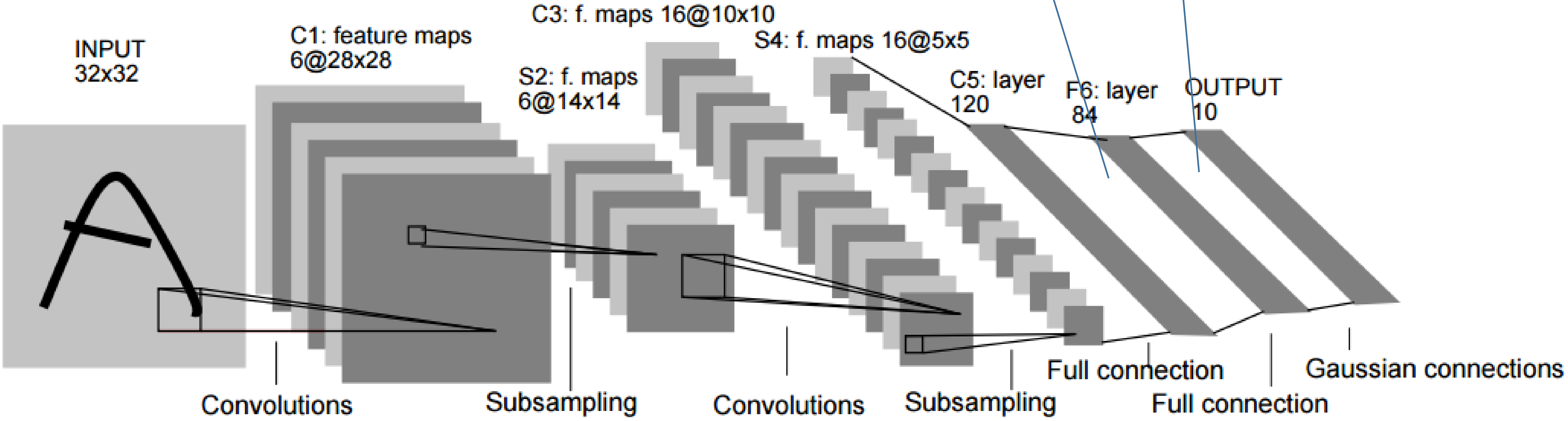


Figure from *Gradient-based learning applied to document recognition, by Y. LeCun, L. Bottou, Y. Bengio and P. Haffner*