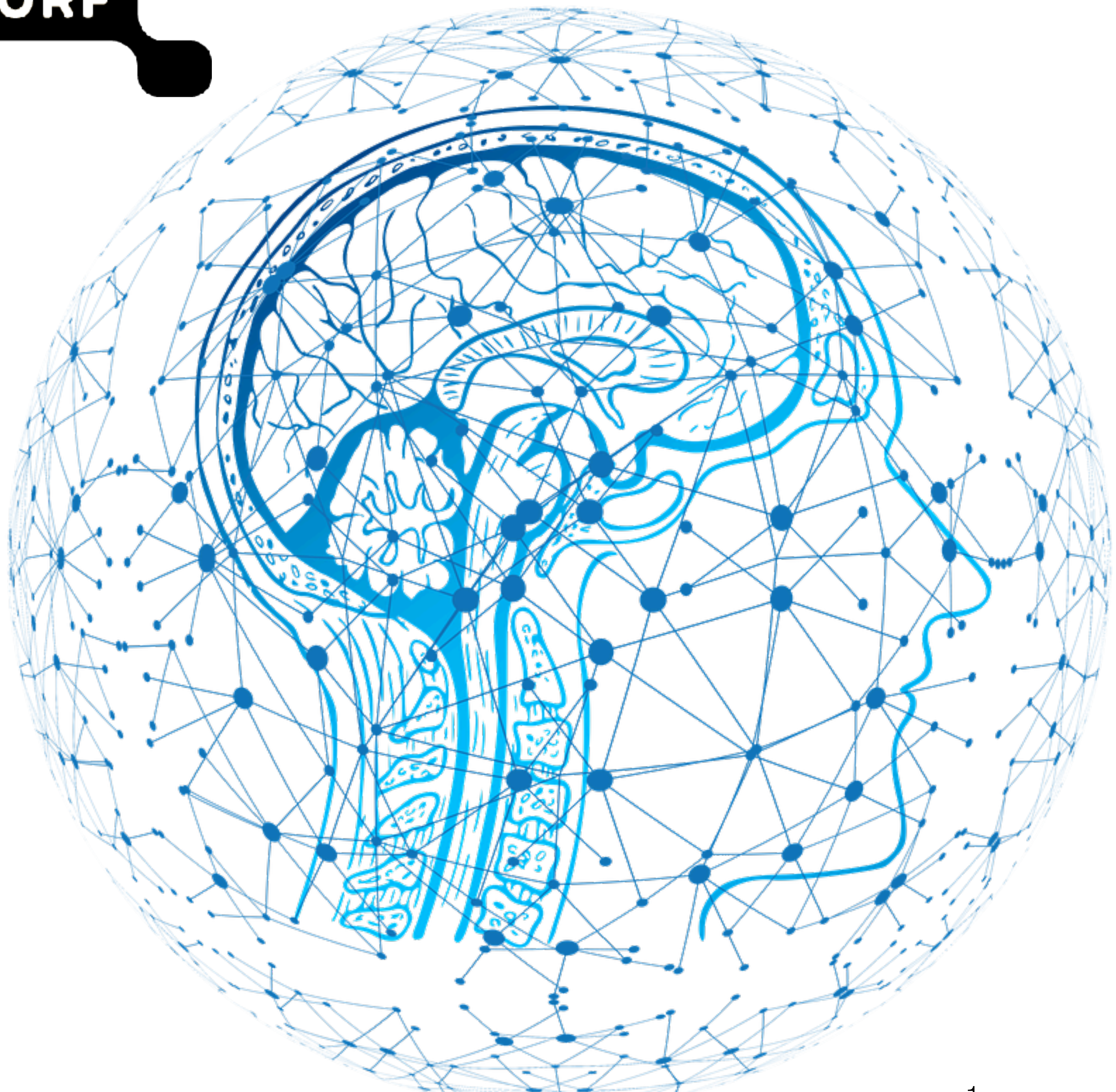


Introduction to Deep Learning

Convolutional Neural Networks

Maxwell Cai

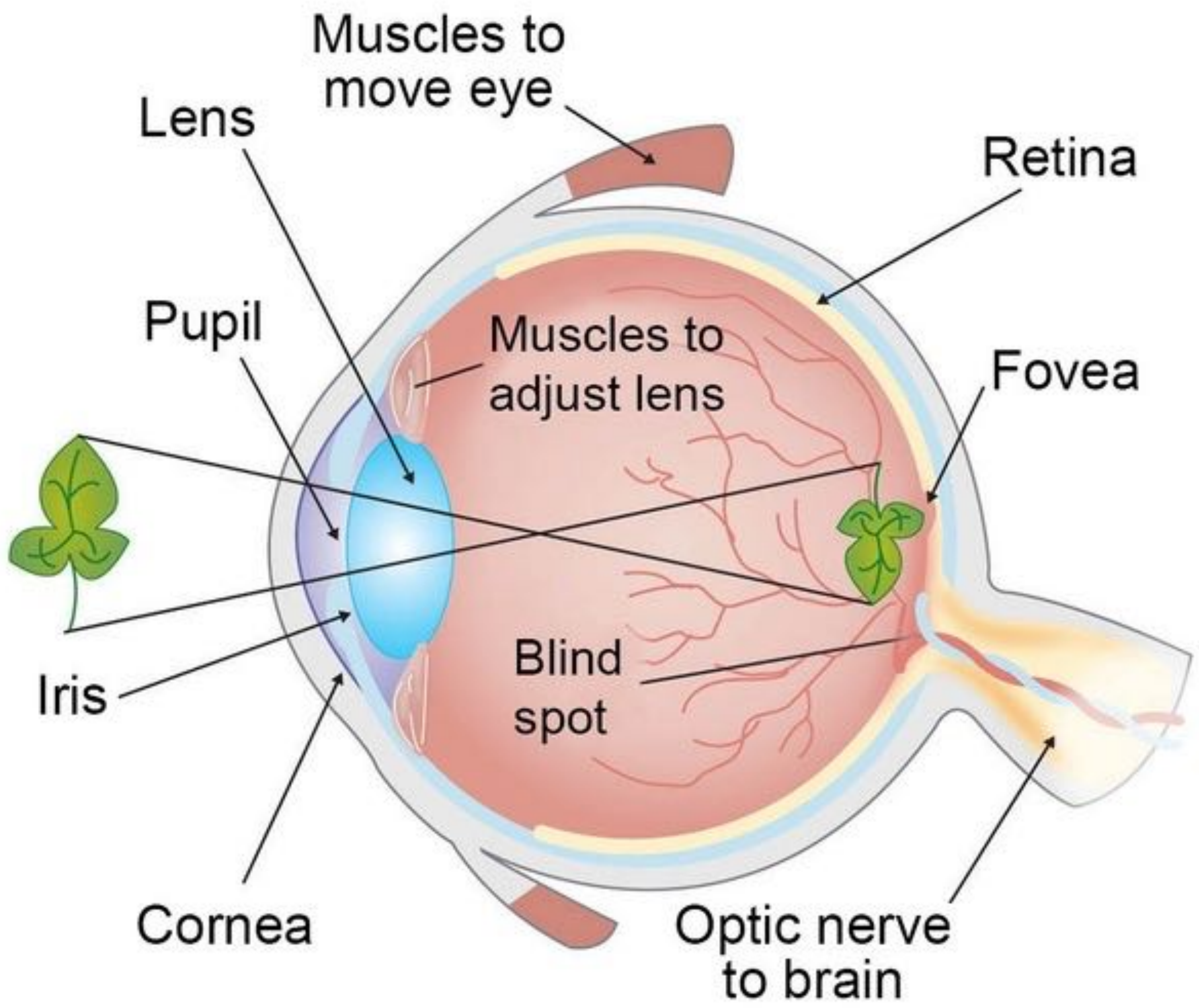
maxwellcai.com

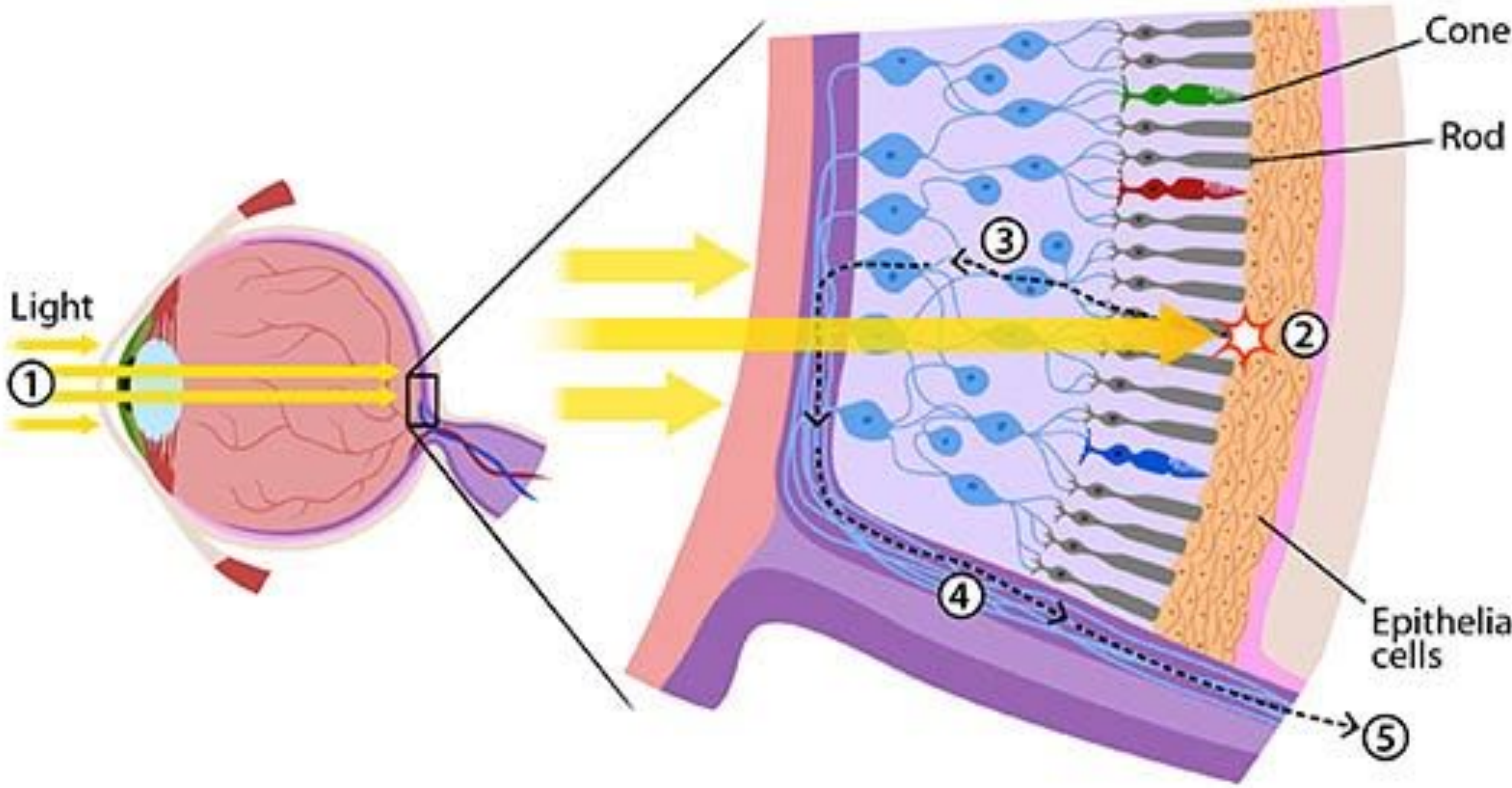
Welcome to the world of computer vision!

How do we let computers “see” something?

Ask ourselves first: How do we (humans) see something?

Human eyes

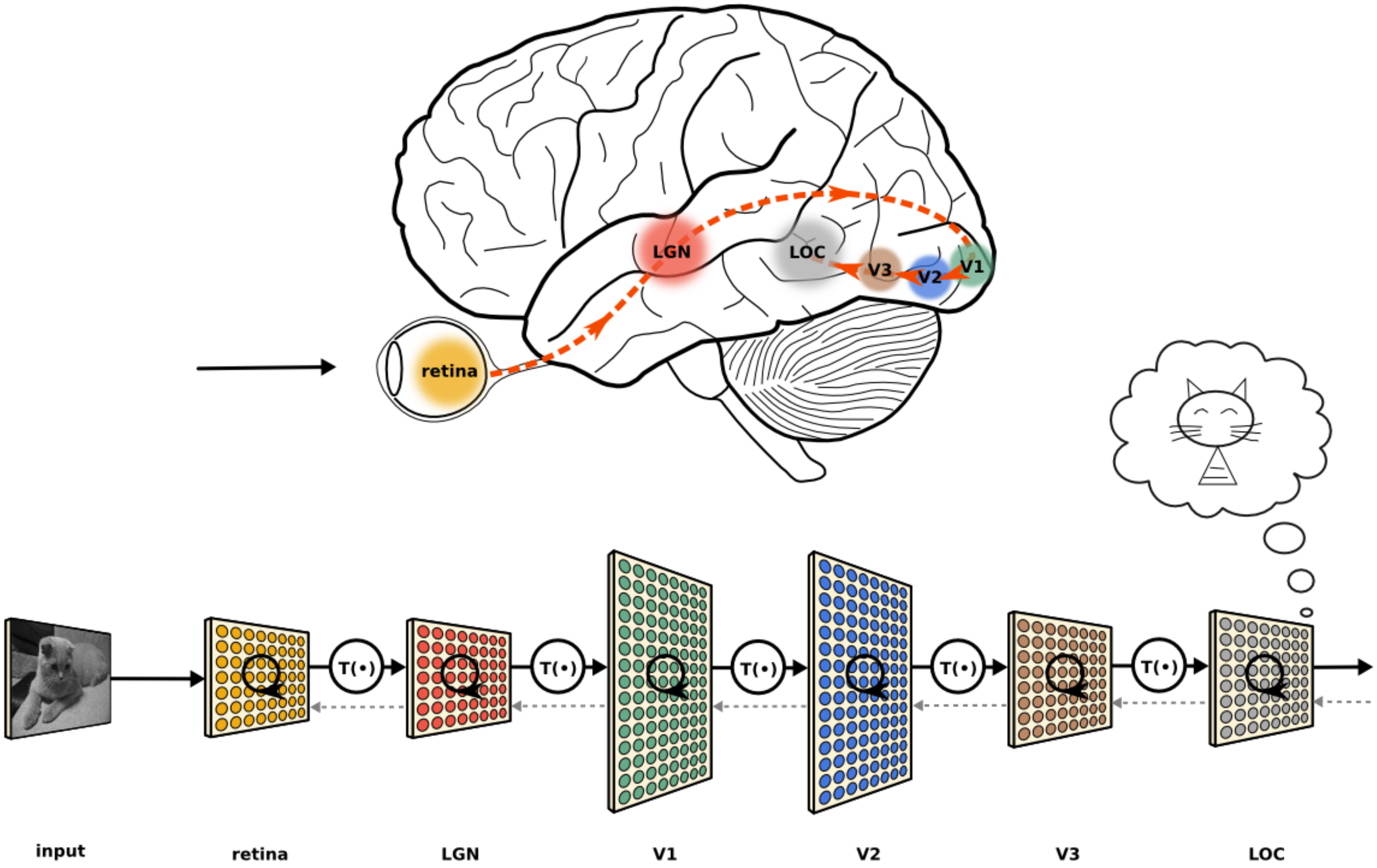




Source: Quora

Human brain: the (real) neural network

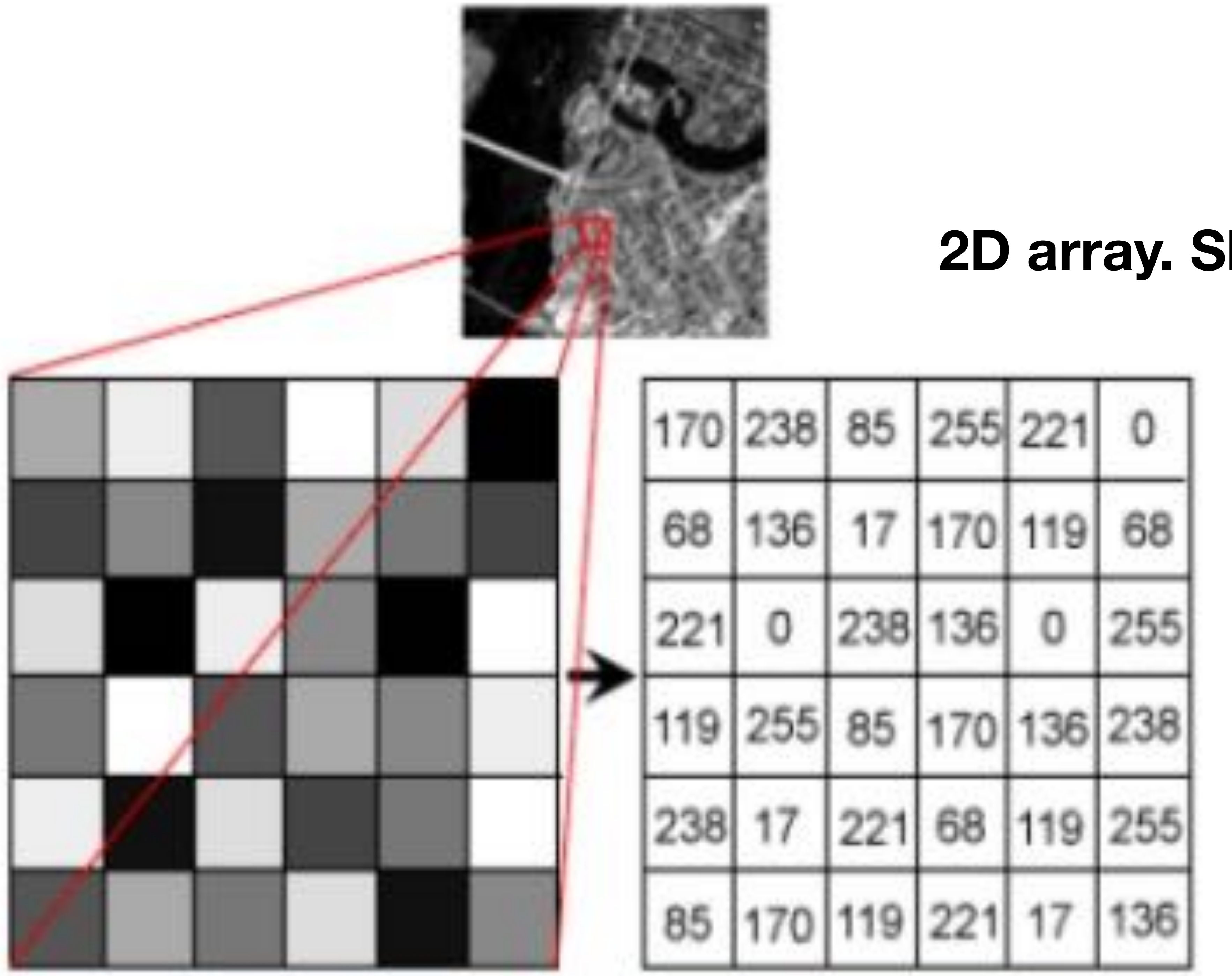
Source: arimaresearch.com



Representation of images in computers

An image

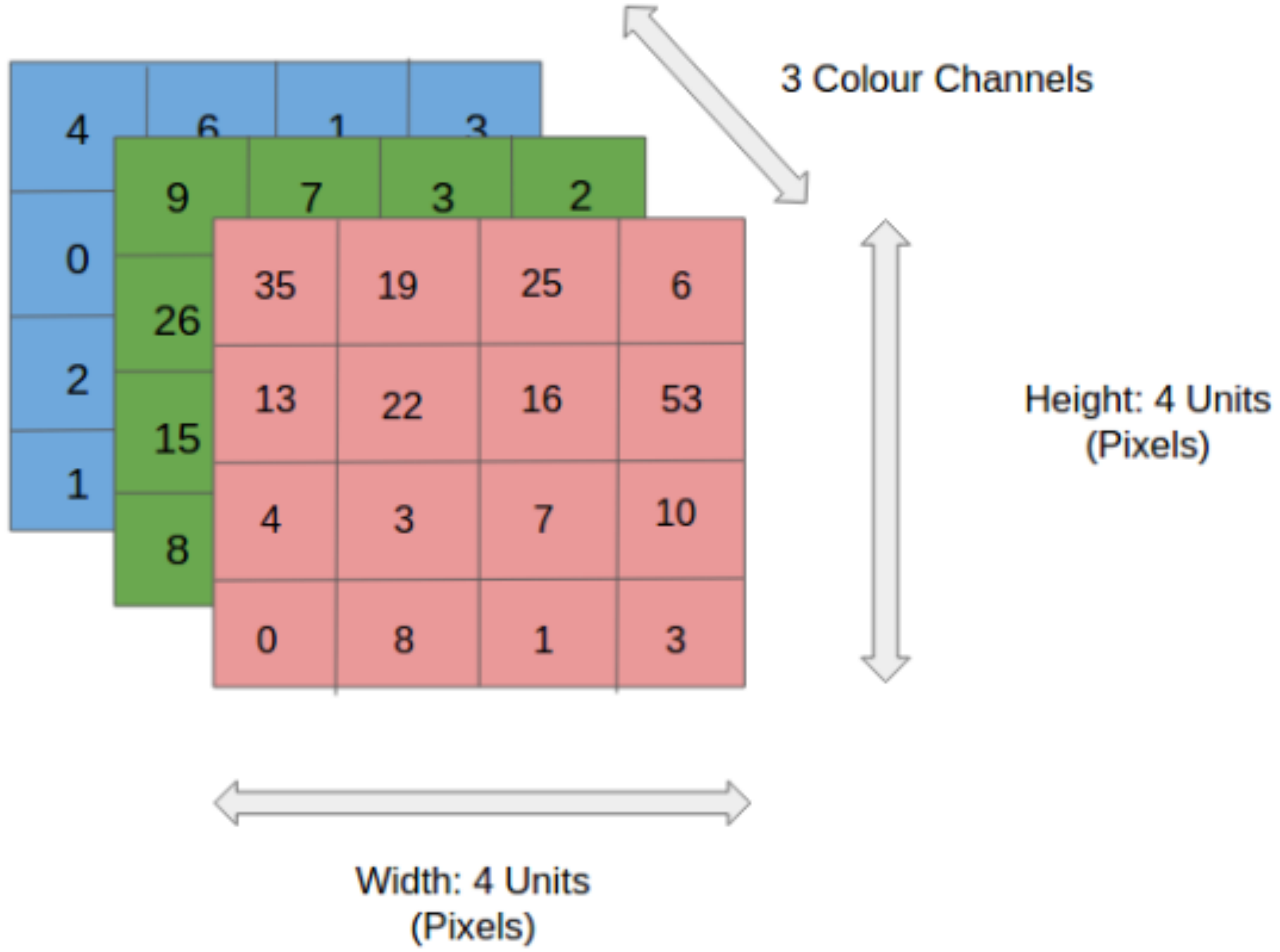
… is a **matrix/array** of intensity values
… usually consists integers of [0, 255] or float points of [0, 1]
… each element of this matrix is called a **pixel**
… can have 1 (greyscale) or multiple (color) **channels**



170	238	85	255	221	0
68	136	17	170	119	68
221	0	238	136	0	255
119	255	85	170	136	238
238	17	221	68	119	255
85	170	119	221	17	136

2D array. Shape: [width, height]

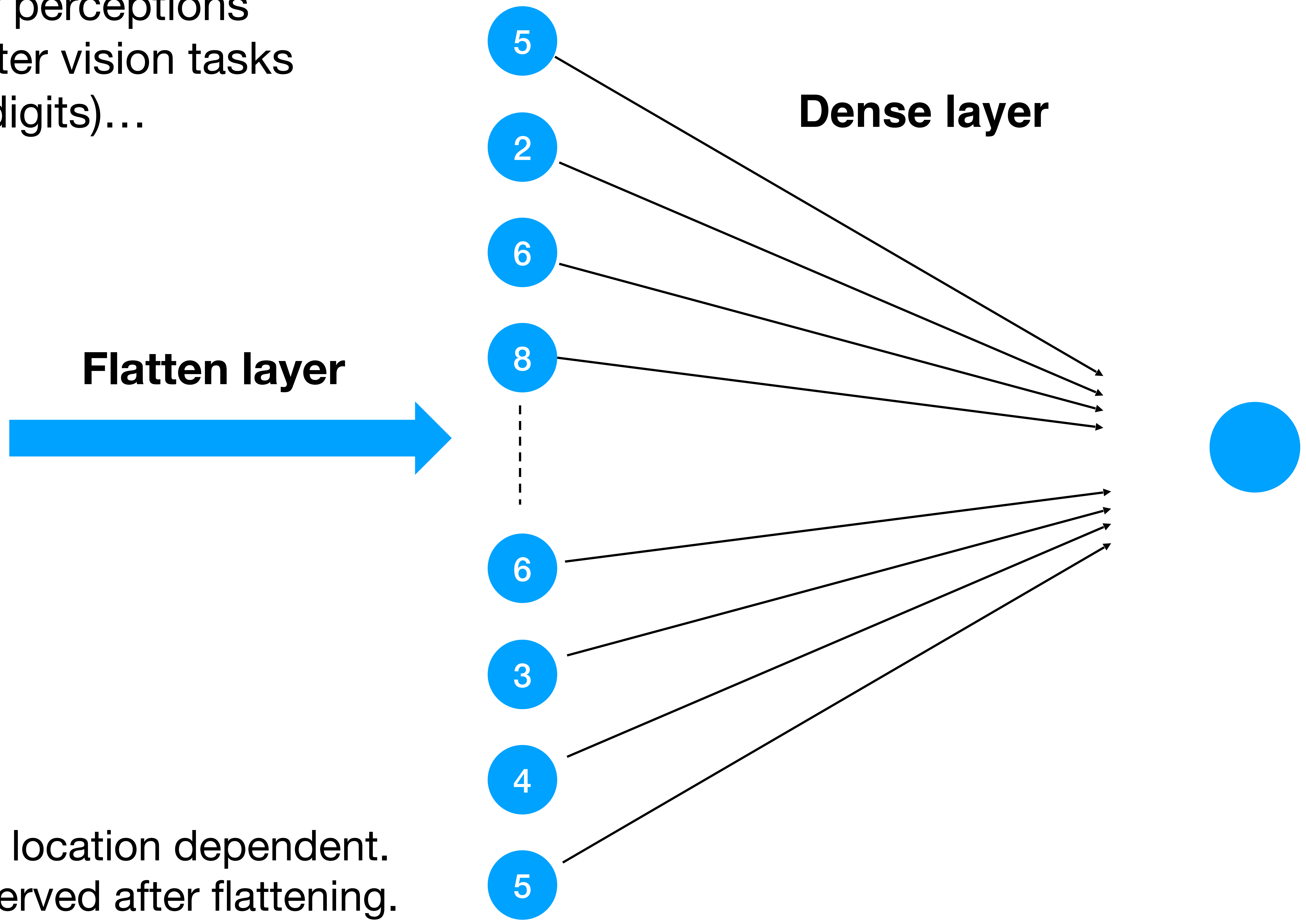
3D array. Shape: [width, height, channel]



Dense network classifier (MLP)

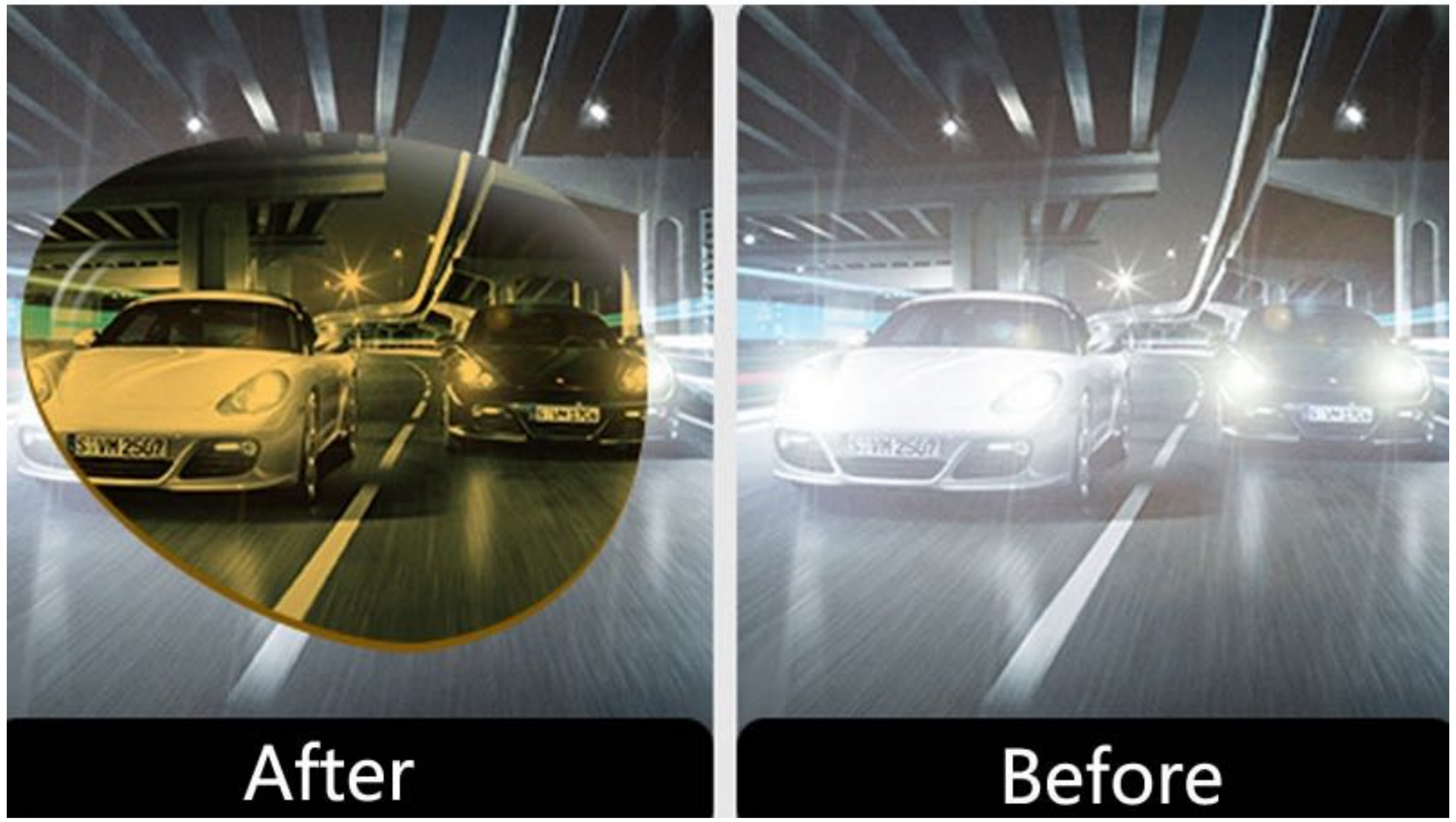
So far, we have used multi-layer perceptions (MLP) to carry out some computer vision tasks (e.g., recognizing hand-written digits)…

5	2	6	8	2	0	1	2
4	3	4	5	1	9	6	3
3	9	2	4	7	7	6	9
1	3	4	6	8	2	2	1
8	4	6	2	3	1	8	8
5	8	9	0	1	0	2	3
9	2	6	6	3	6	2	1
9	8	8	2	6	3	4	5



Problem:
- The resulting encoding is pixel location dependent.
- Spatial relationship is not preserved after flattening.

Filters in our daily lives



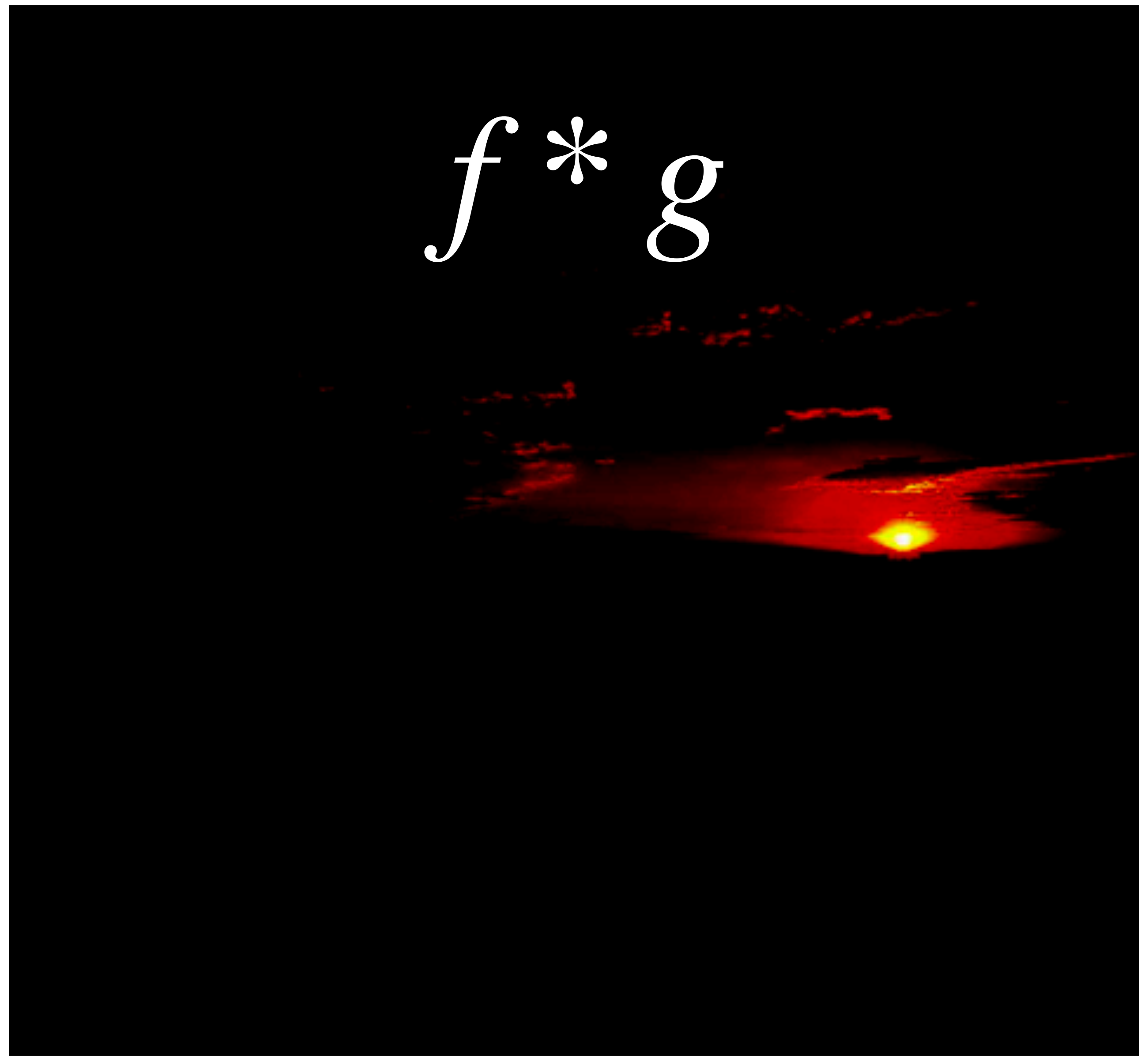
Forget about the commercial advertisements themselves…
Just think about the mathematical principles behind them…

Filter in image processing

f

$+$

Filter g $=$

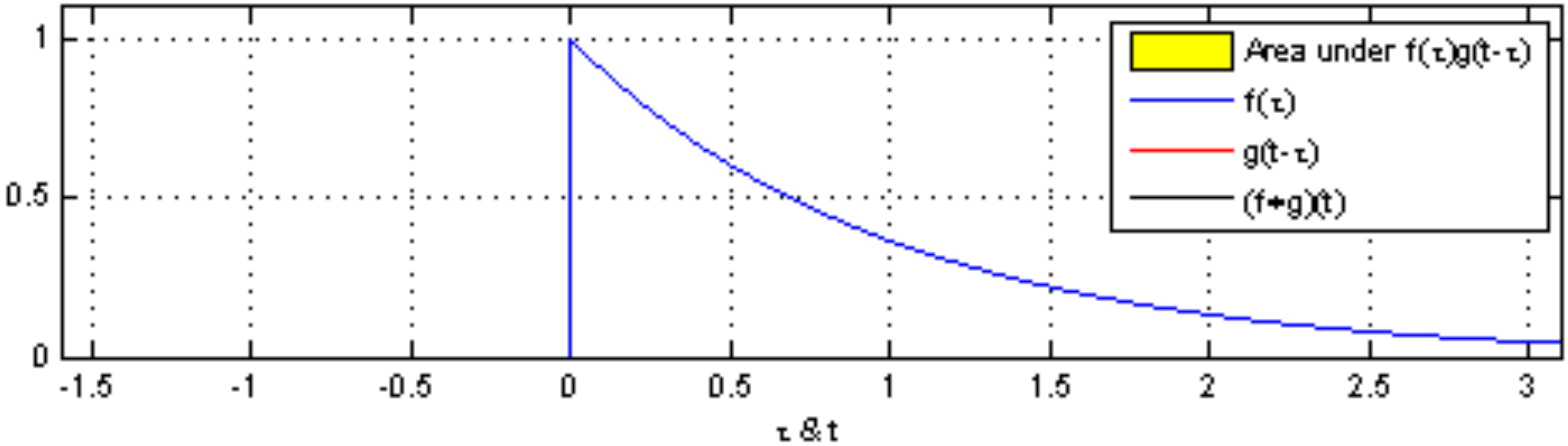


Filter h $=$

Using **different filters**, we can see the **same signal** in **different perspectives**.

Convolution

Typically, a filter applies a **convolution** operation upon the original signal.



Function: $f(t)$

Kernel: $g(t)$

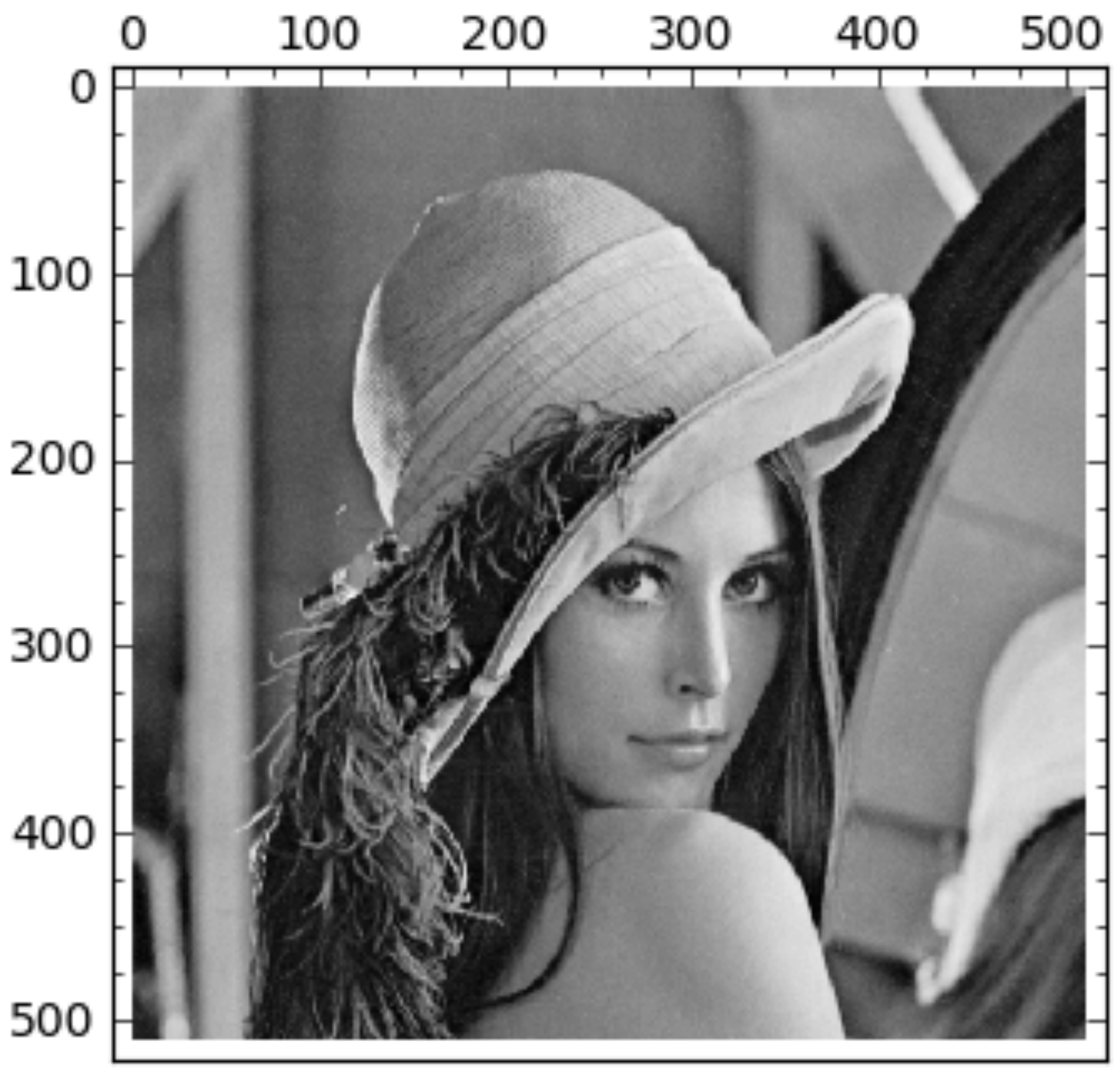
Convolution: $(f * g)(t)$

No signal → no response

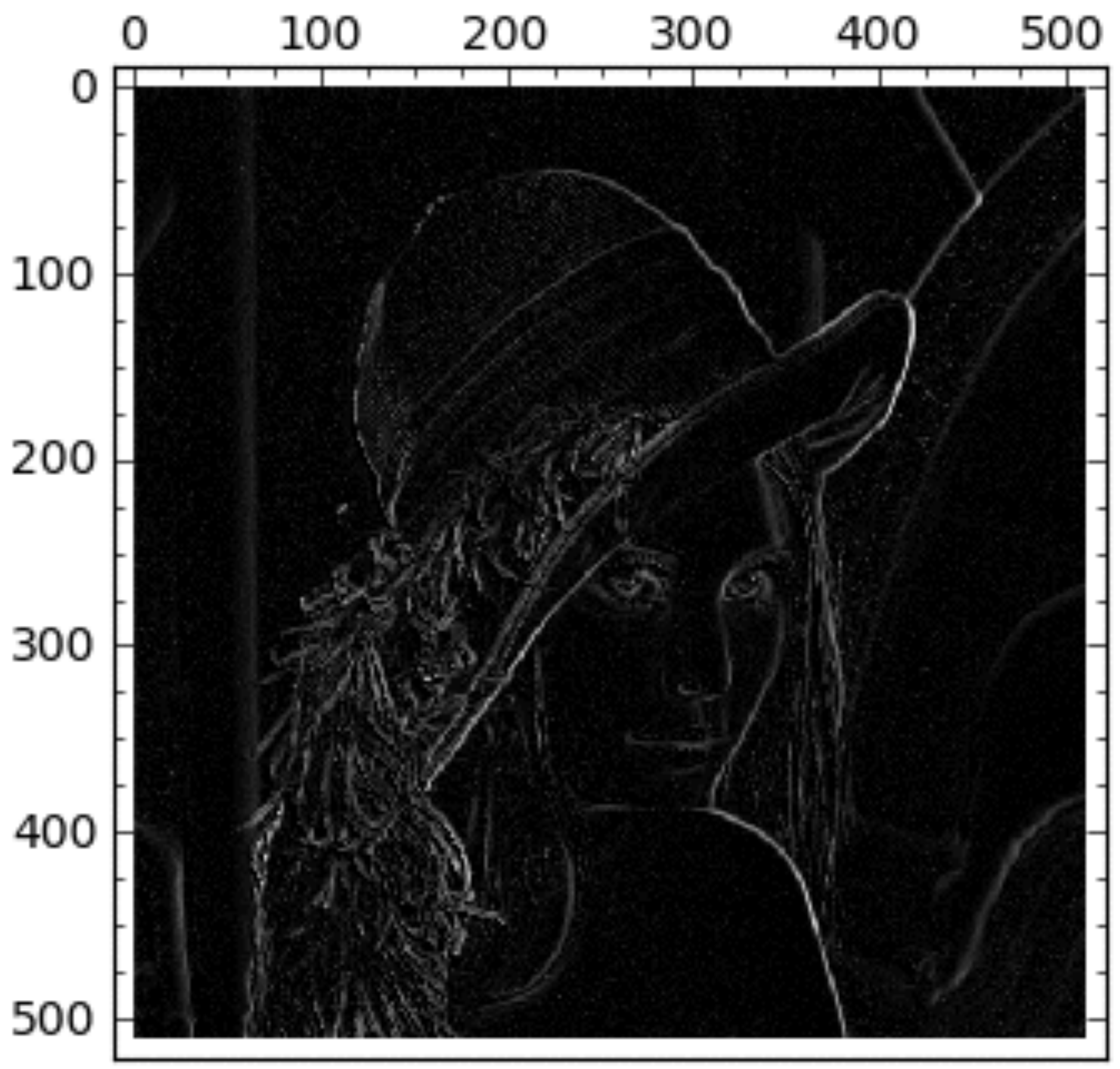
Strong signal → strong response

Filtering: a signal processing technique

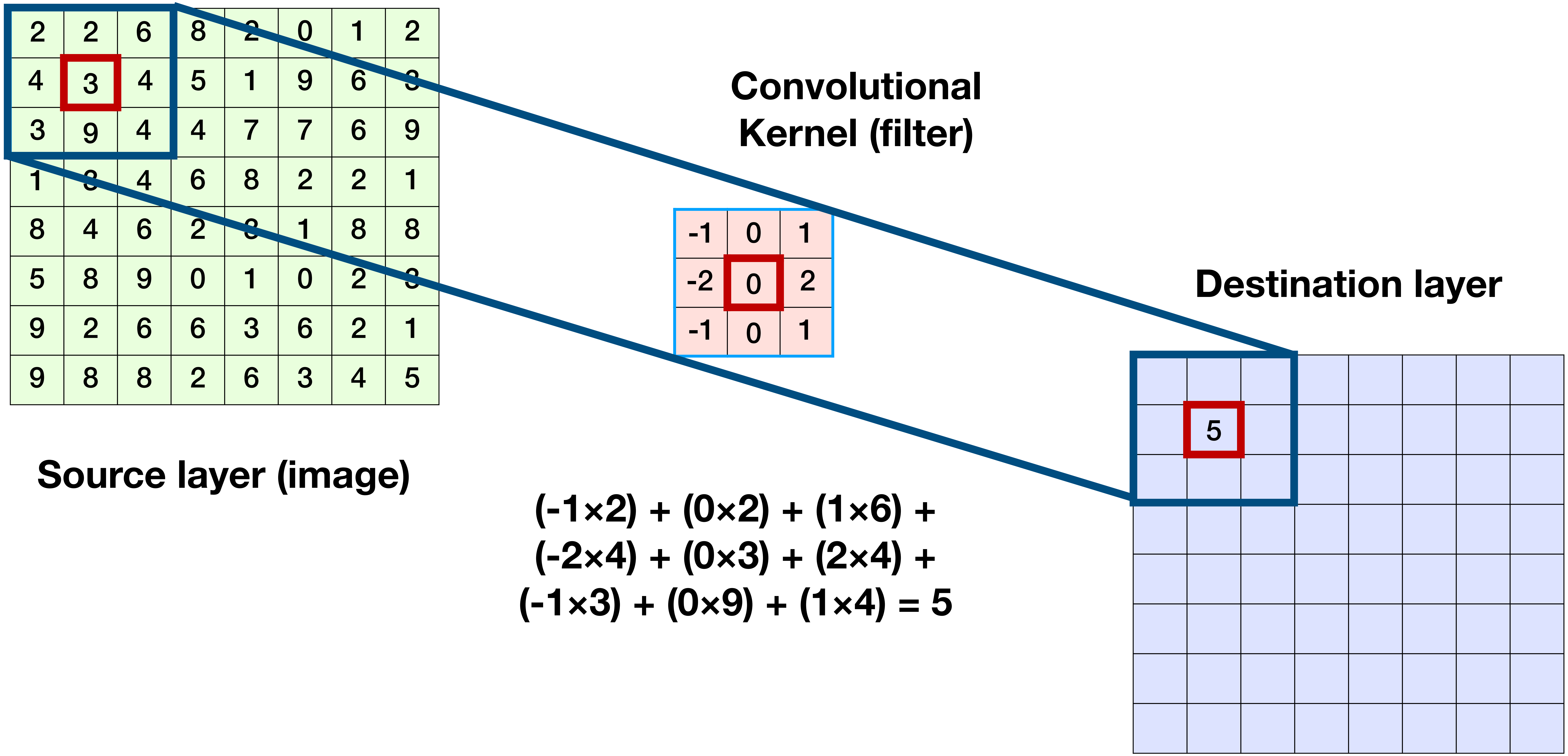
Convolution



Convolution

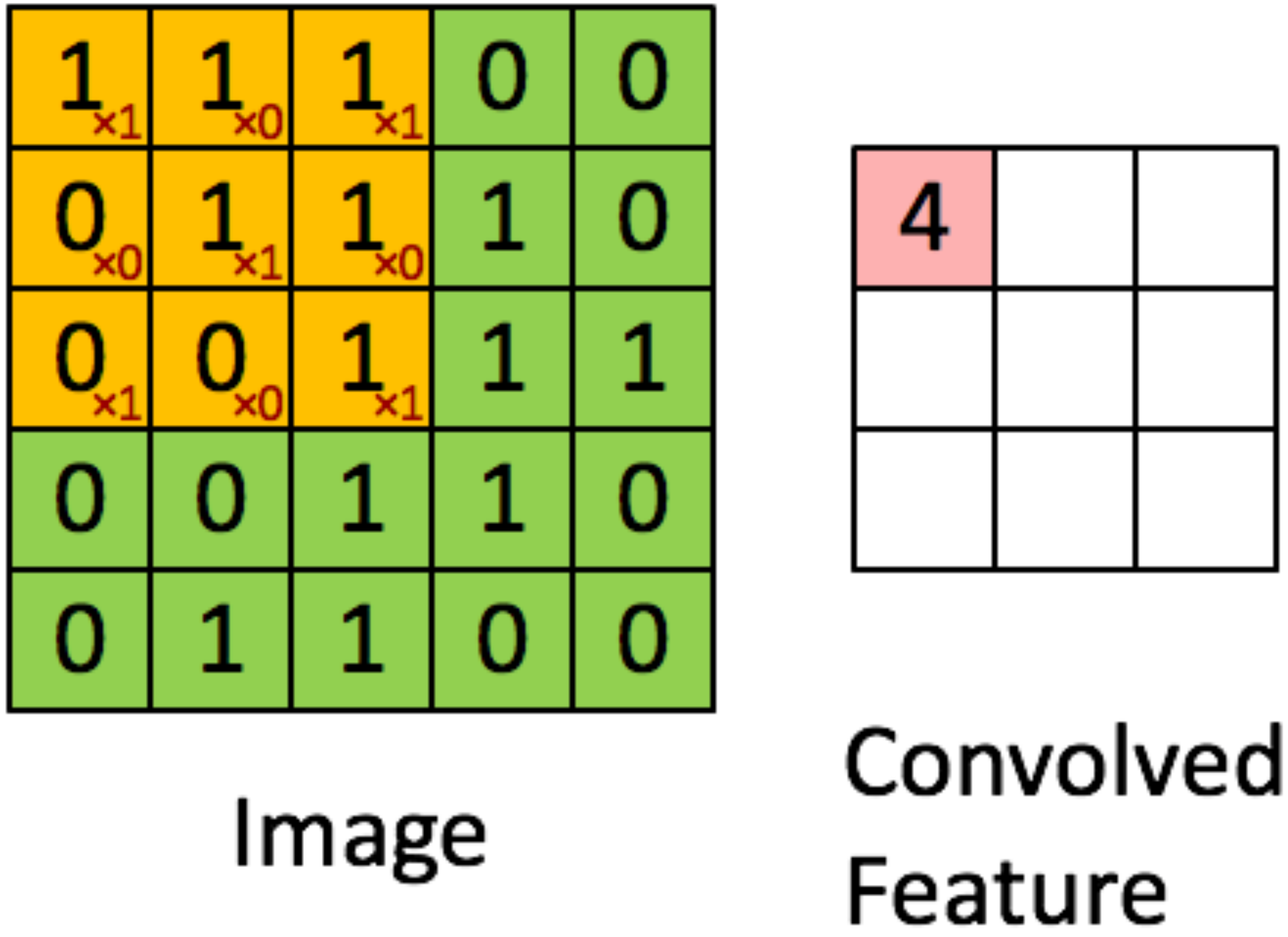


Convolution

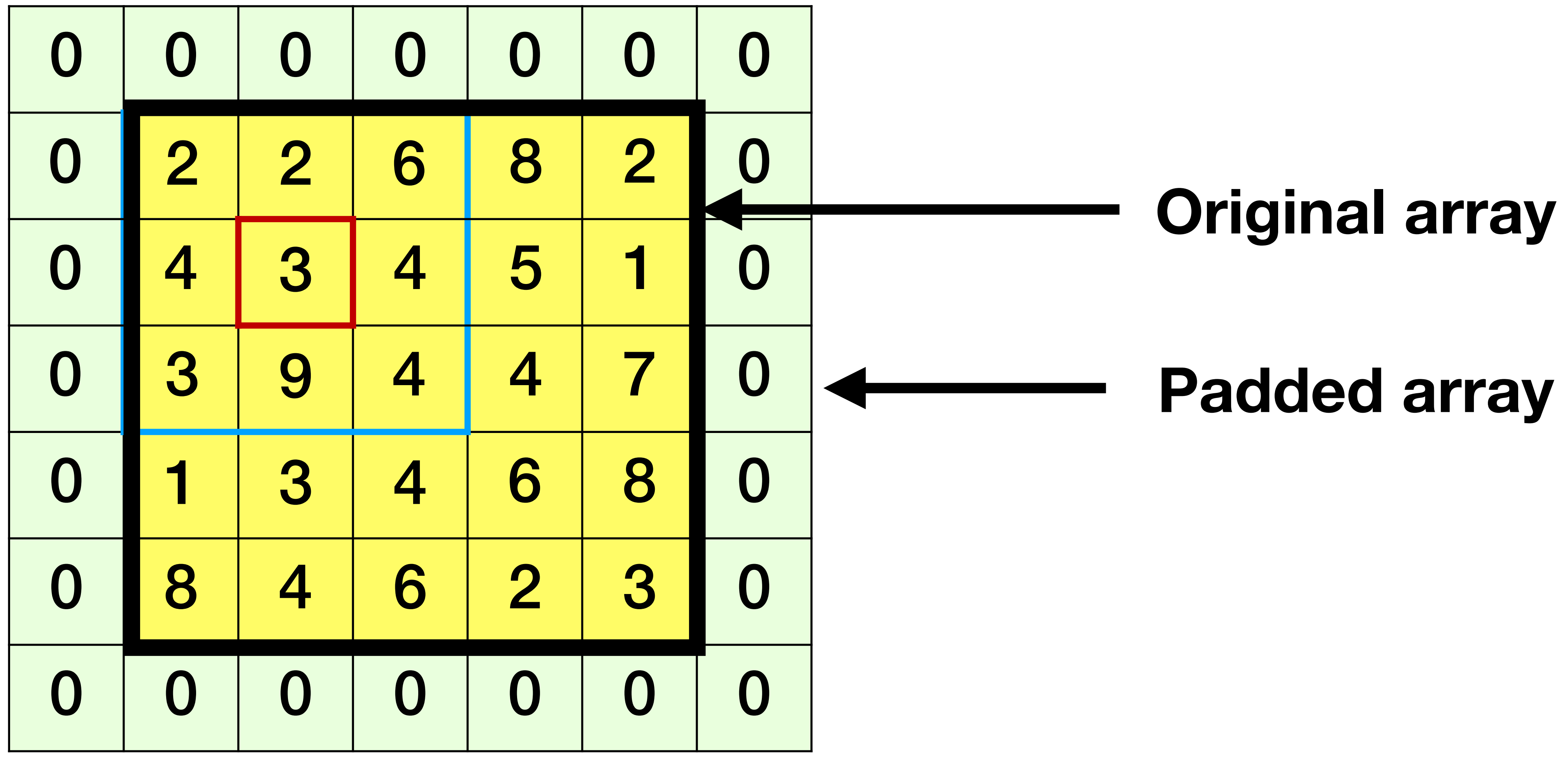


$$(-1\times 2) + (0\times 2) + (1\times 6) +$$
$$(-2\times 4) + (0\times 3) + (2\times 4) +$$
$$(-1\times 3) + (0\times 9) + (1\times 4) = 5$$

Convolution

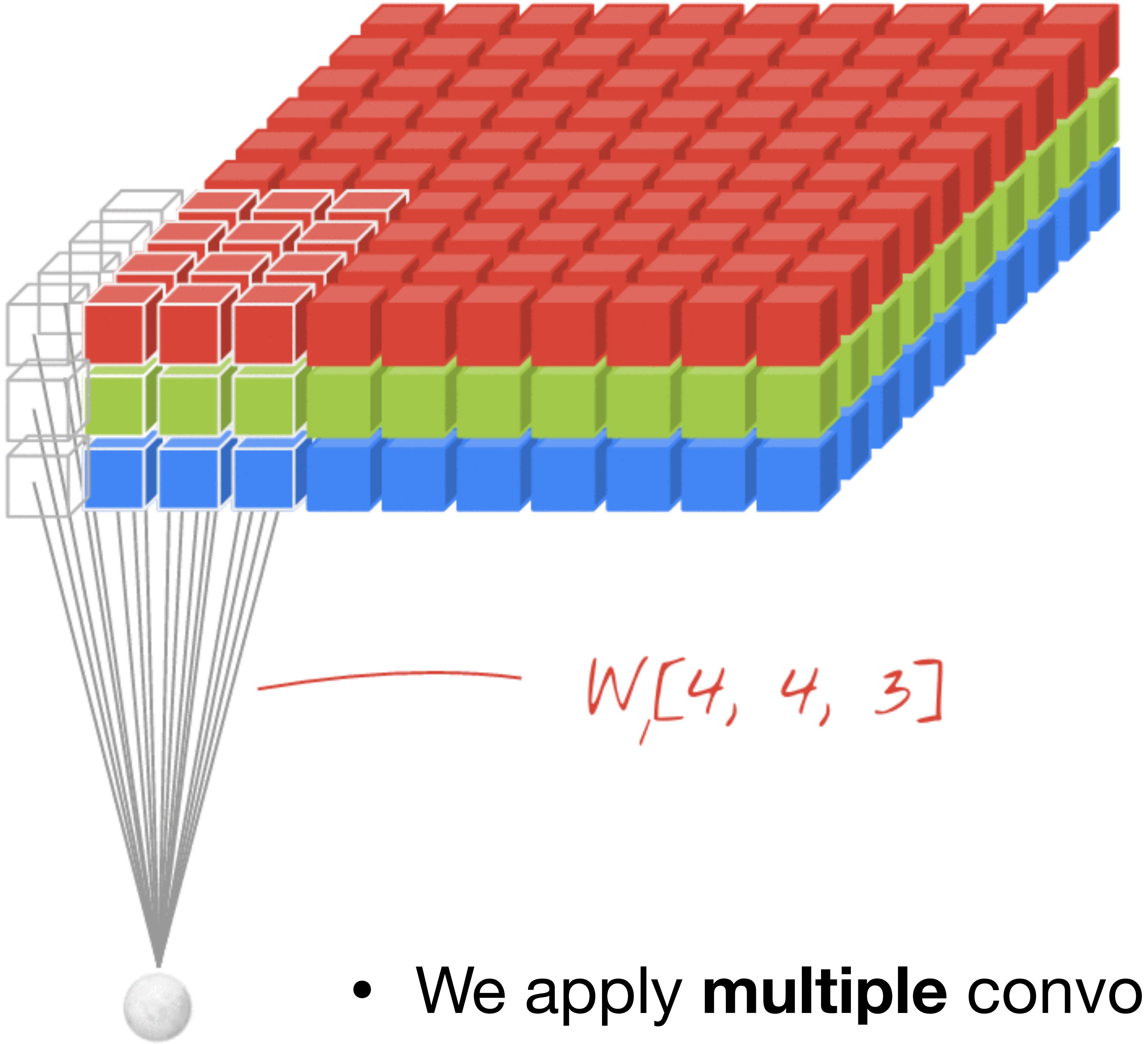


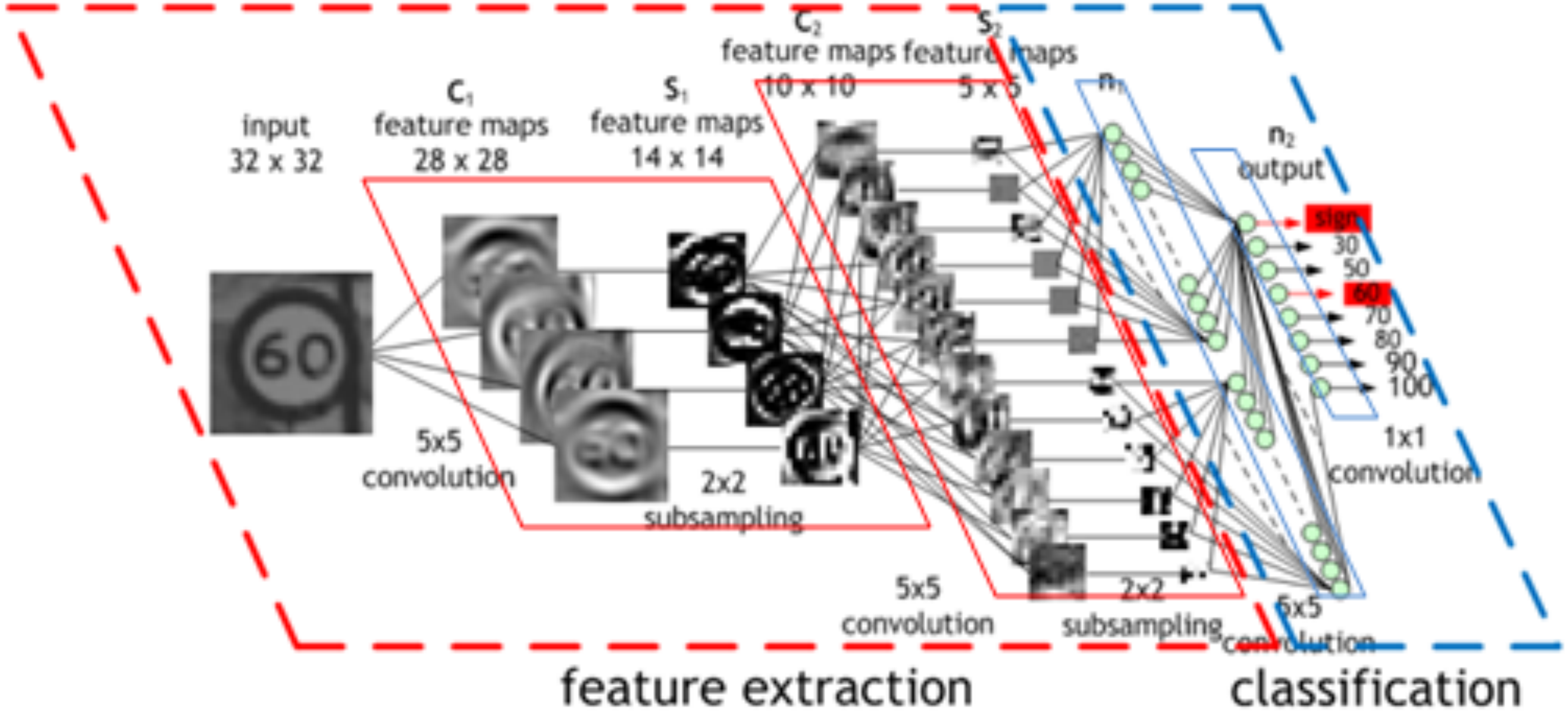
- The kernel is **shifted** over the image with a step size, and computes the output for each position.
- The step size is called **stride**.
- Output has **smaller** dimension than input:
  dim(output) = dim(input) - (dim(kernel) - 1)
- **Padding** is used to solve this problem, which artificially make the image "bigger" by adding synthesis data (typically 0-padding)



Animation: Sumit Saha/Towards Data Science

Multiple convolutional channels



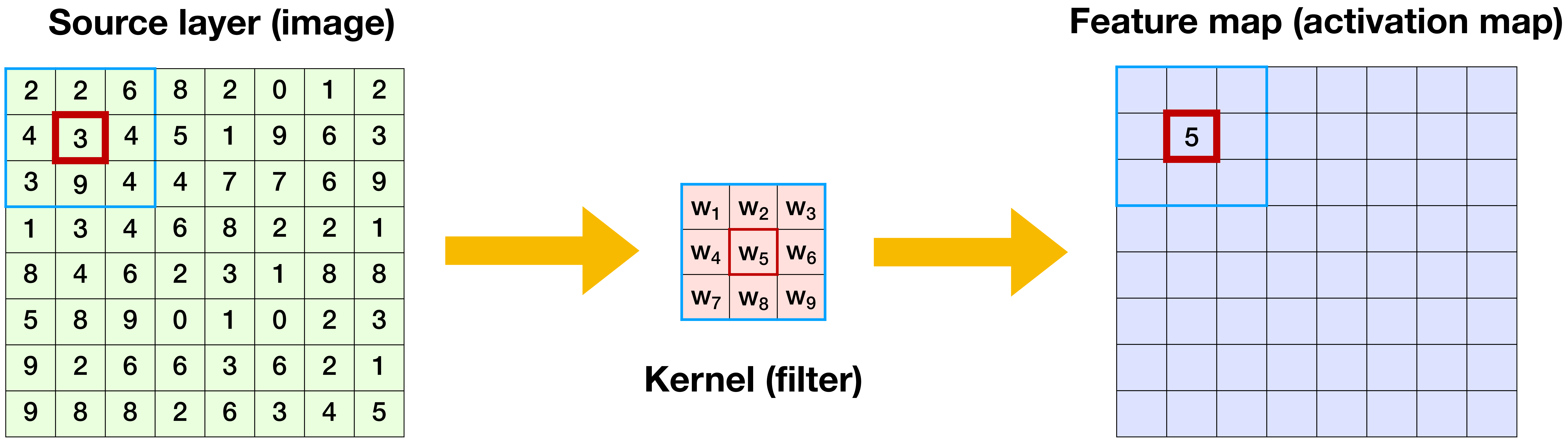


- We apply **multiple** convolutional **filters/kernels** to the **same** image.
- Each filter results in one convolutional **channel**.
- By learning the same image from different channels, one can detect complex patterns.

Animation: Martin Görner

Image: Maurice Peemen

Automatic Kernel Determination

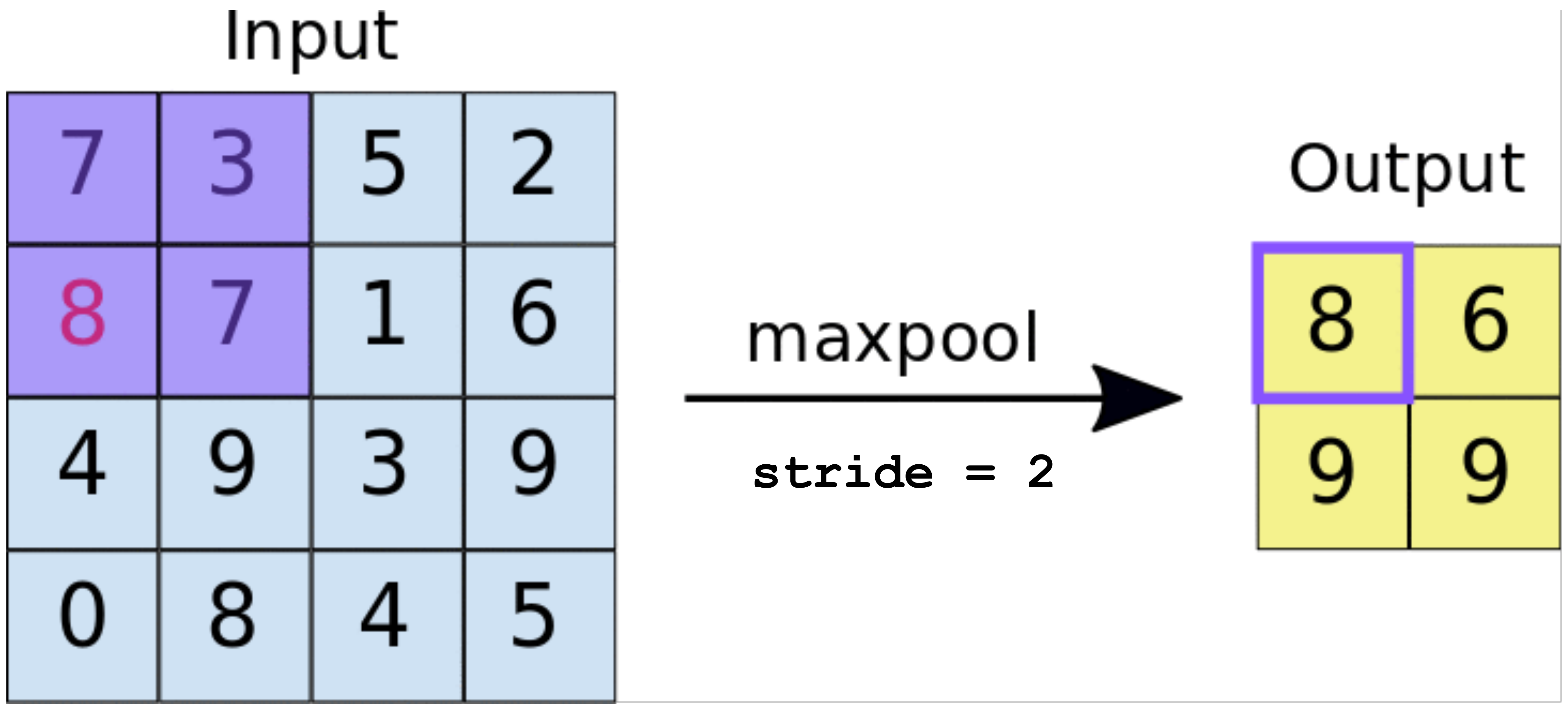


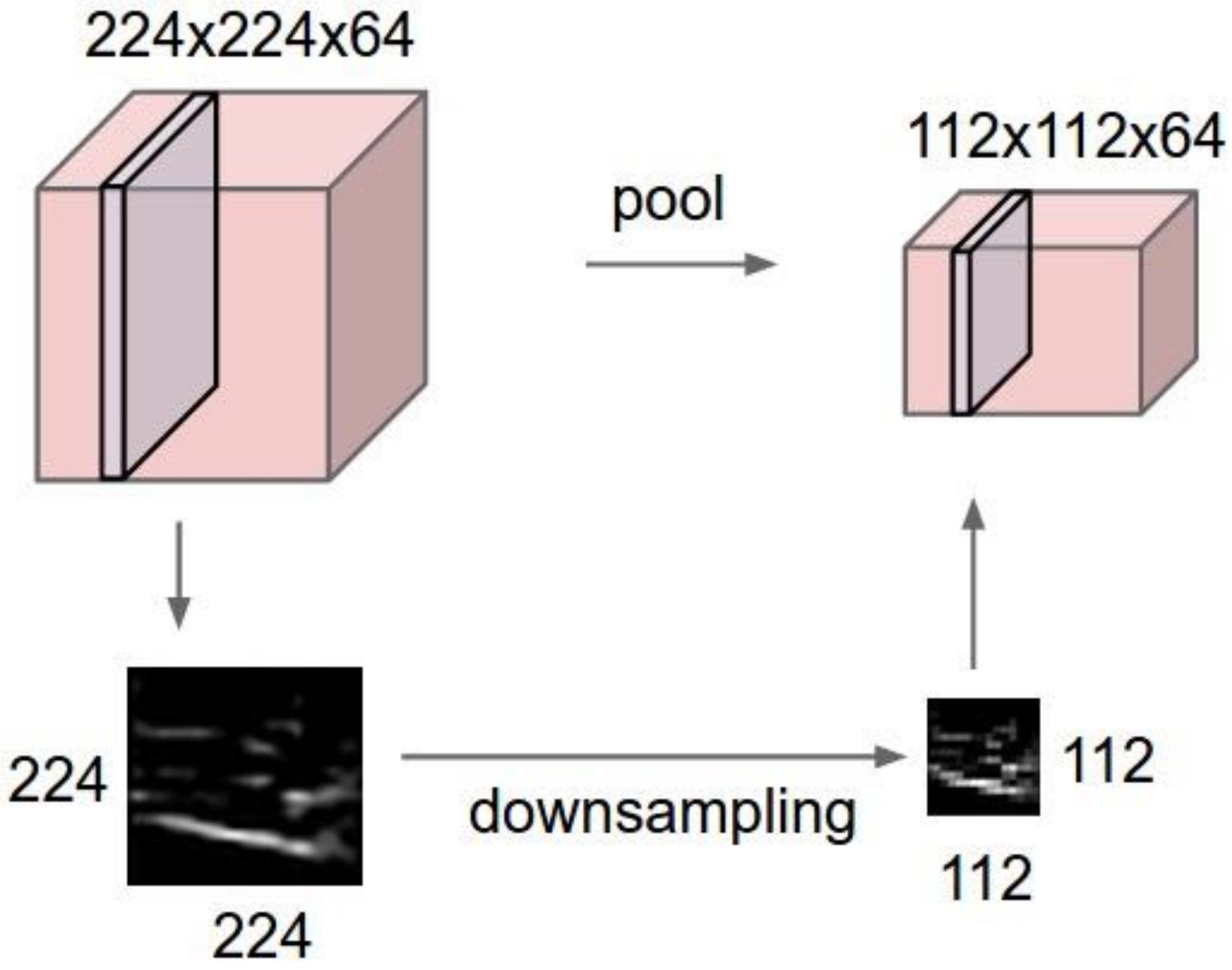
How do we know which kernels to use?

- In **traditional** signal processing, a filter is designed by human experts.
- In **deep learning**, data are too complex and too many different kernels are needed.
- Therefore, kernels are no longer fixed. They are initialized **randomly**.
- Kernels are updated through **backward propagation**, in the way that it **learns** which features to detect.
- **Gradient descent** algorithms are used.

Pooling (downsampling)

We need to **discard** information **gradually**.
Pooling is a way of information **abstraction**.
Pooling is usually applied **after** convolutions.

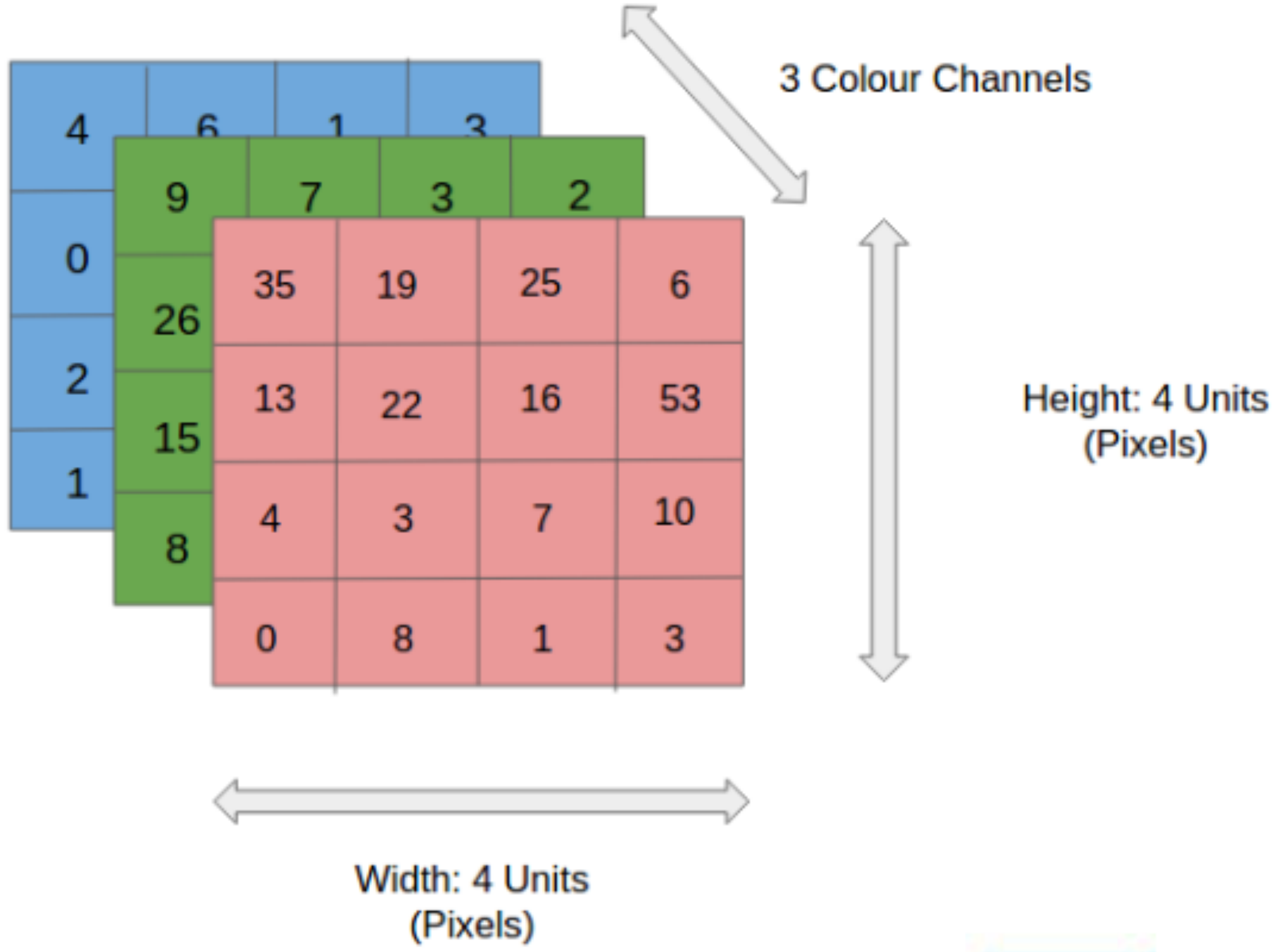


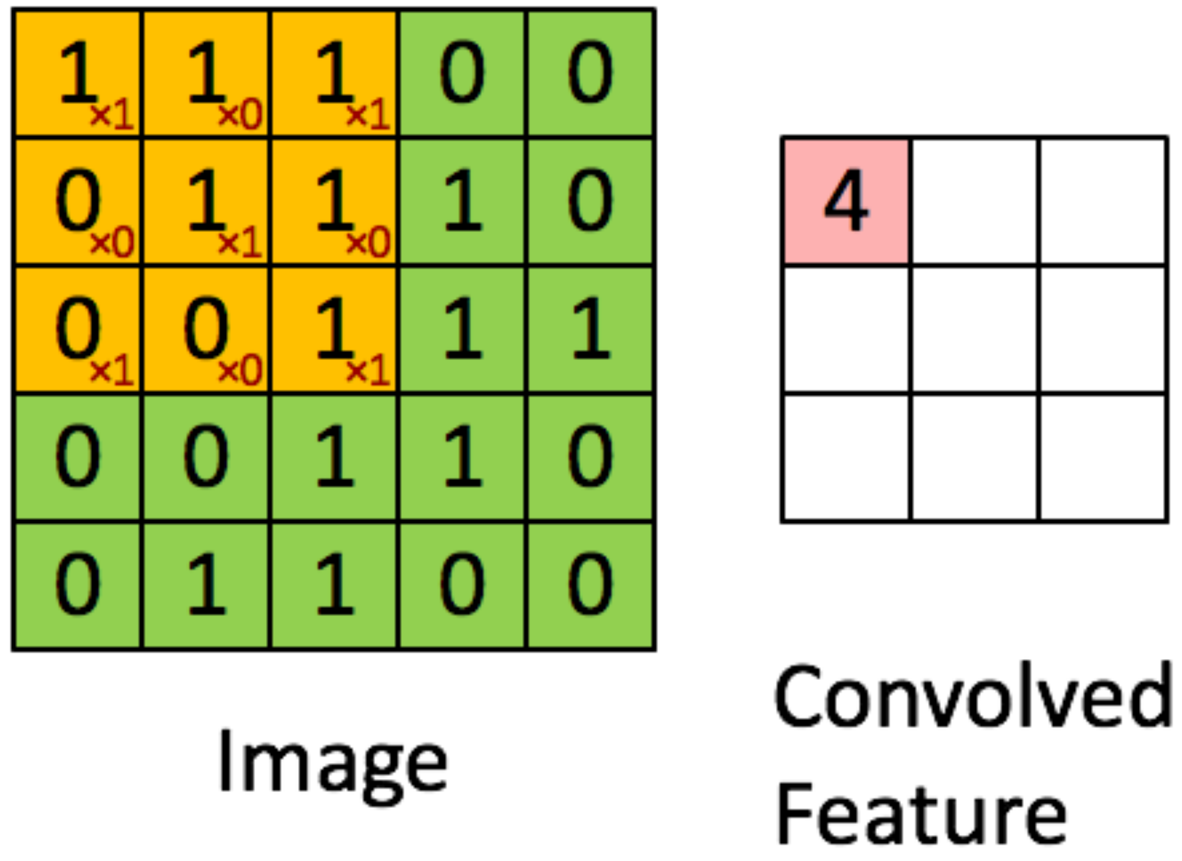


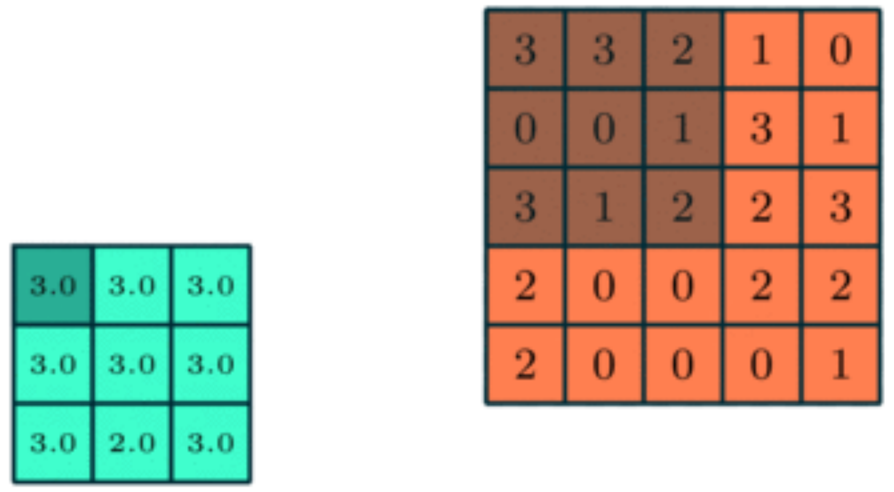
Max pooling: keep the **strongest** signal
Average pooling: use the **local average** as the signal

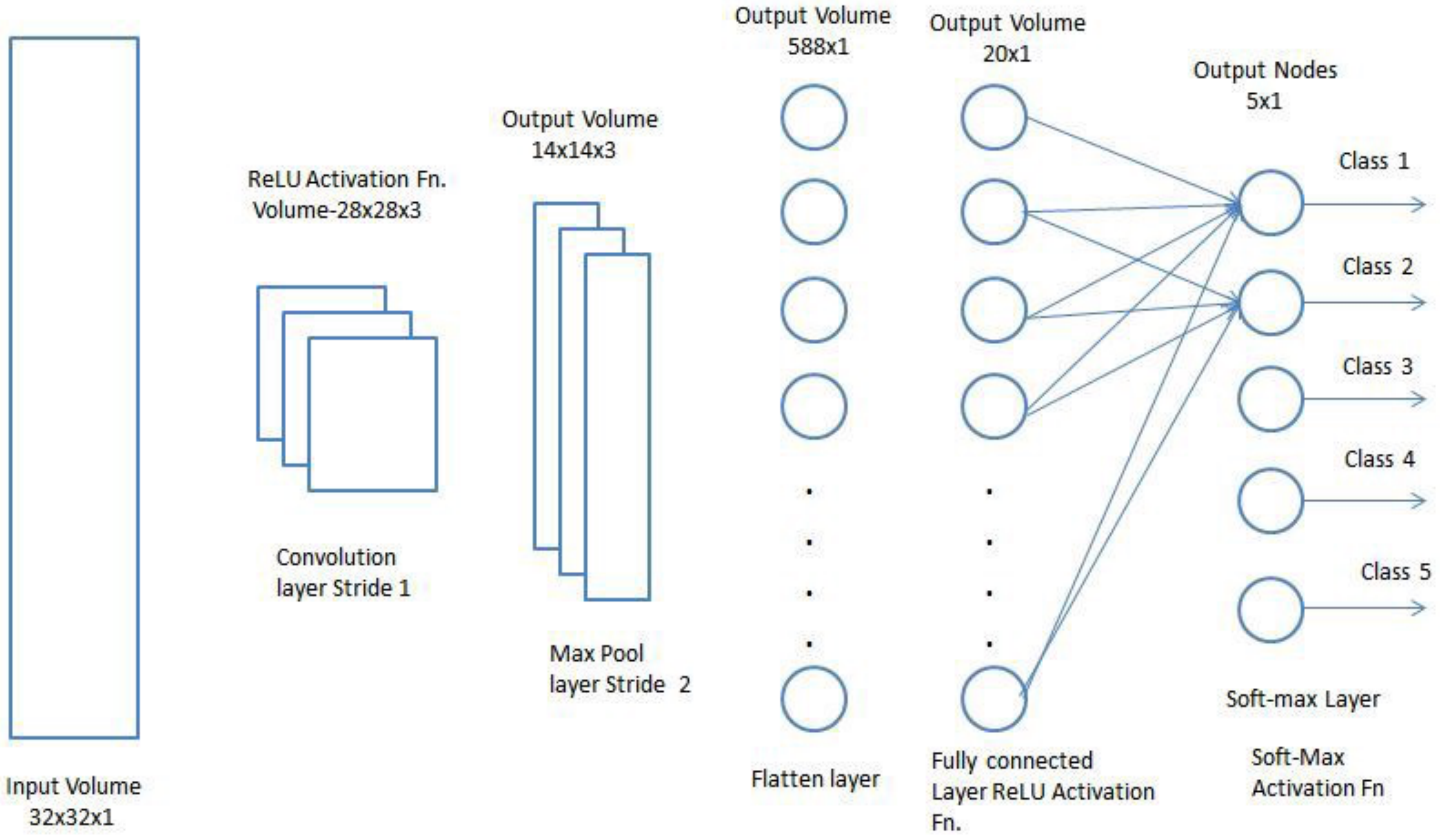
Animation: Google Developers

CNN Architecture



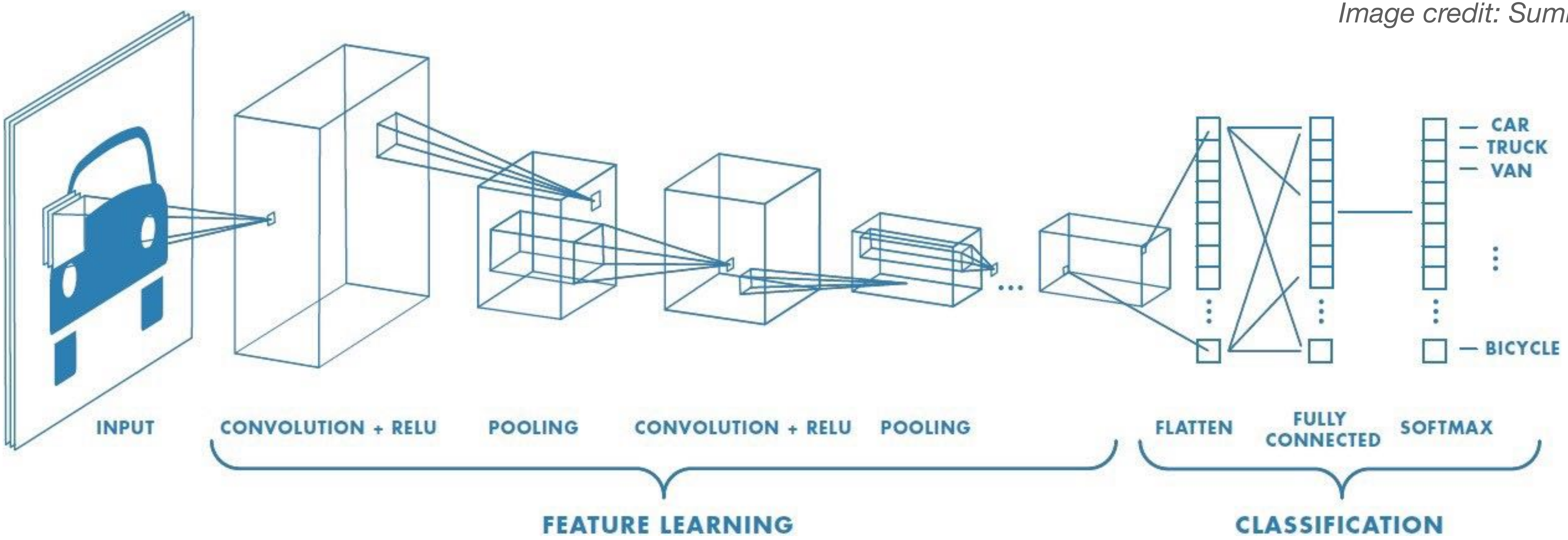






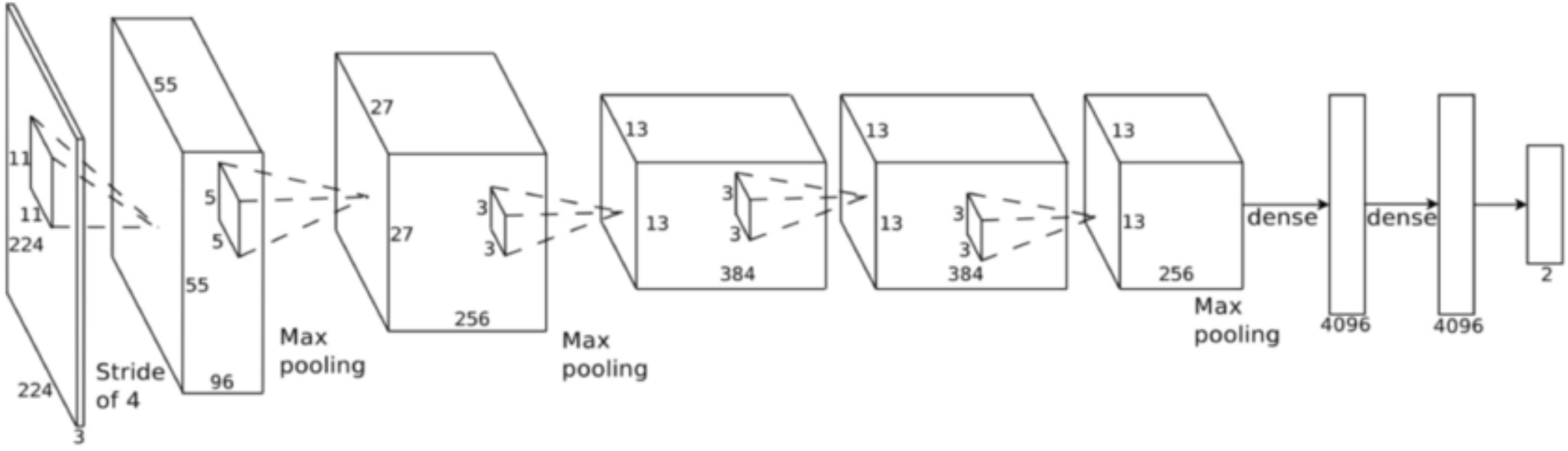
Convolution Neural Network (CNN)

Image credit: Sumit Saha



- Typically, **images** become increasingly **smaller** as they go to the **deeper** layers;
- Typically, the number of **filters increases** in the **deeper** layers;
  - Low-level features are limited, thus requires few filters;
  - High-level features are rich, thus requires many filters;
- The **decision making** (i.e., classification) is made in the **last layers**, typically with **dense** layers and the `softmax` activation function.

AlexNet



Krizhevsky et al. 2012 (NIPS). *58k+ citations!*

Do we need to go deeper?

A modern CNN uses **multiple convolution modules.**
“We need to go deeper” — a popular view in the AI community before 2015. Really?

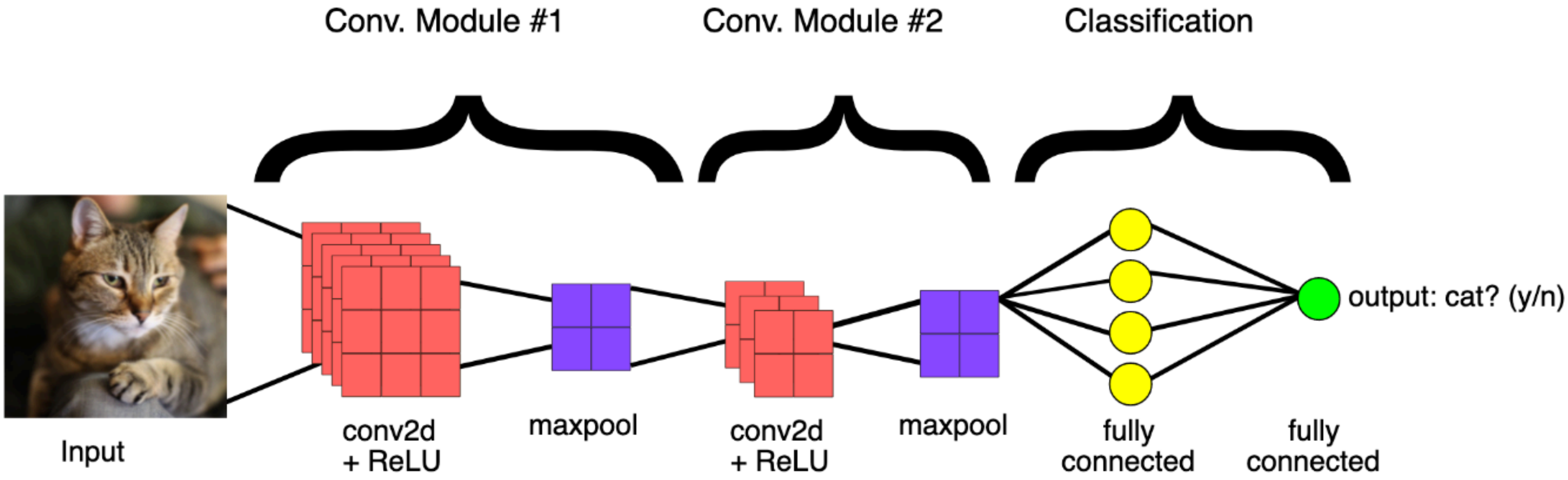


Image credit: Google Developers

Do we need to go deeper?

A modern CNN uses **multiple convolution modules.**
“We need to go deeper” — a popular view in the AI community before 2015. Really?

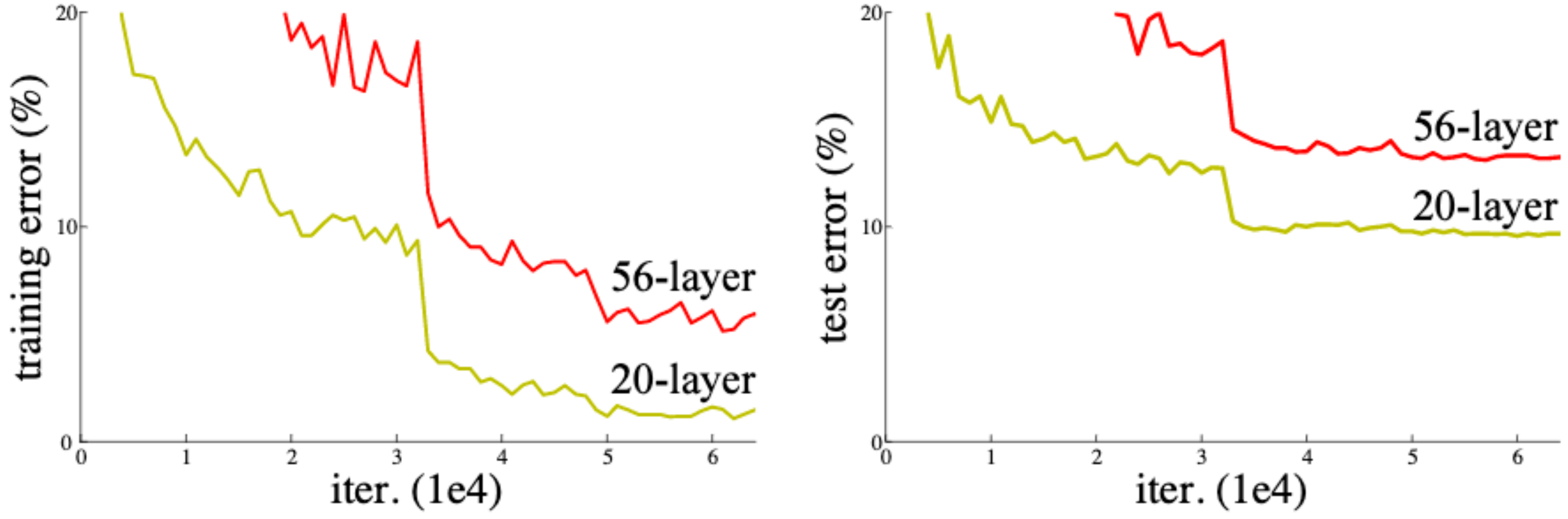


Figure 1. Training error (left) and test error (right) on CIFAR-10 with 20-layer and 56-layer “plain” networks. The deeper network has higher training error, and thus test error. Similar phenomena on ImageNet is presented in Fig. 4.

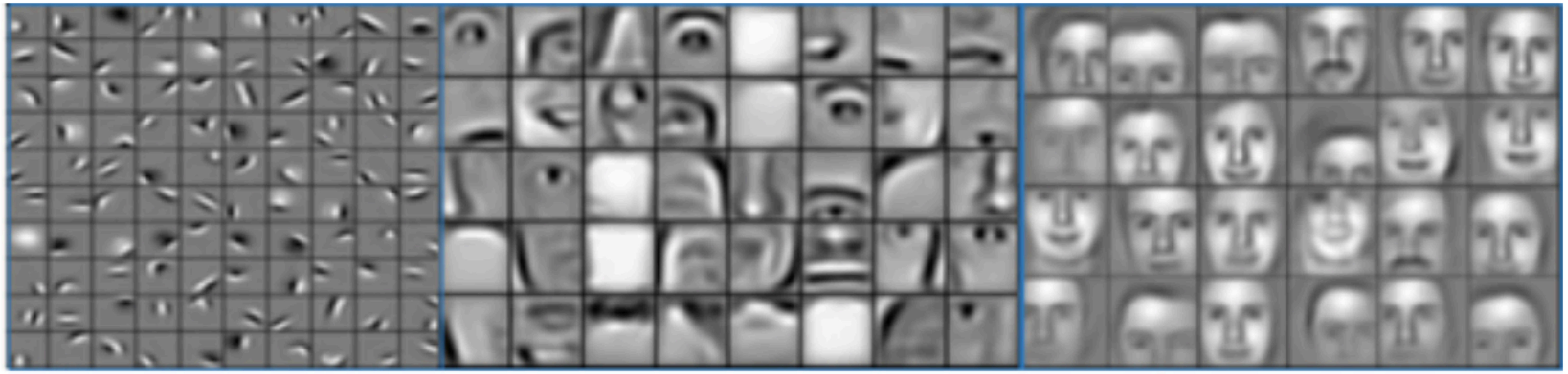
Answer: yes and no.

Deeper networks have larger training/testing error, mainly due to the vanishing gradient problem.

But deeper networks are needed to deal with more complex data.

He et al. (2015, arXiv:1512.03385)

Residual blocks



Low-level features

High-level features

Swallow layer

Deep layer

Specific

Abstract

Shortcut

How about letting deep layers to have direct access to low-level features?

ResNet

He et al. (2015, arXiv:1512.03385)

Transfer Learning

Recall that the convolutional kernels are **learned** during the training process through backward propagation.
The trained kernels contains **knowledge** to detect patterns.
Why not let a new CNN to **inherit these knowledge**?

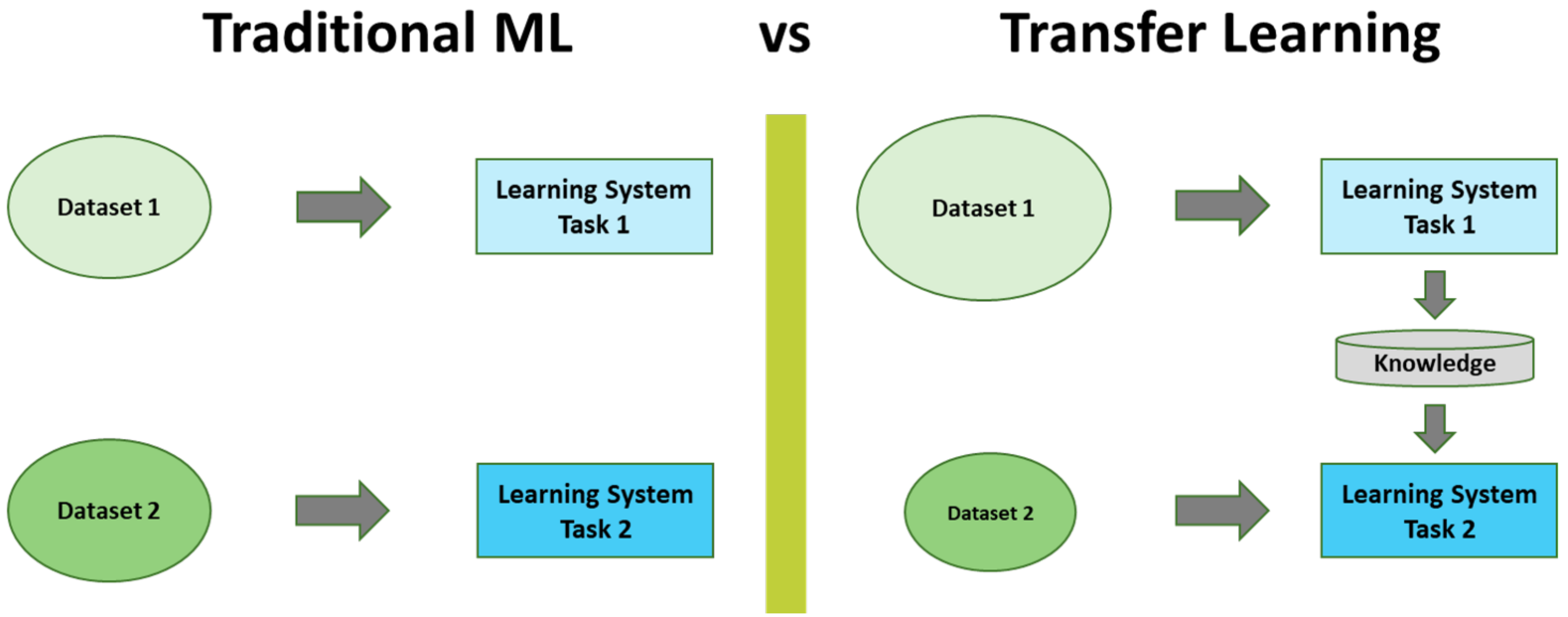


Image source: Dipanjan Sarkar/TowardsDataScience

Transfer Learning

Recall that the convolutional kernels are **learned** during the training process through backward propagation.
The trained kernels contains **knowledge** to detect patterns.
Why not let a new CNN to **inherit these knowledge**?

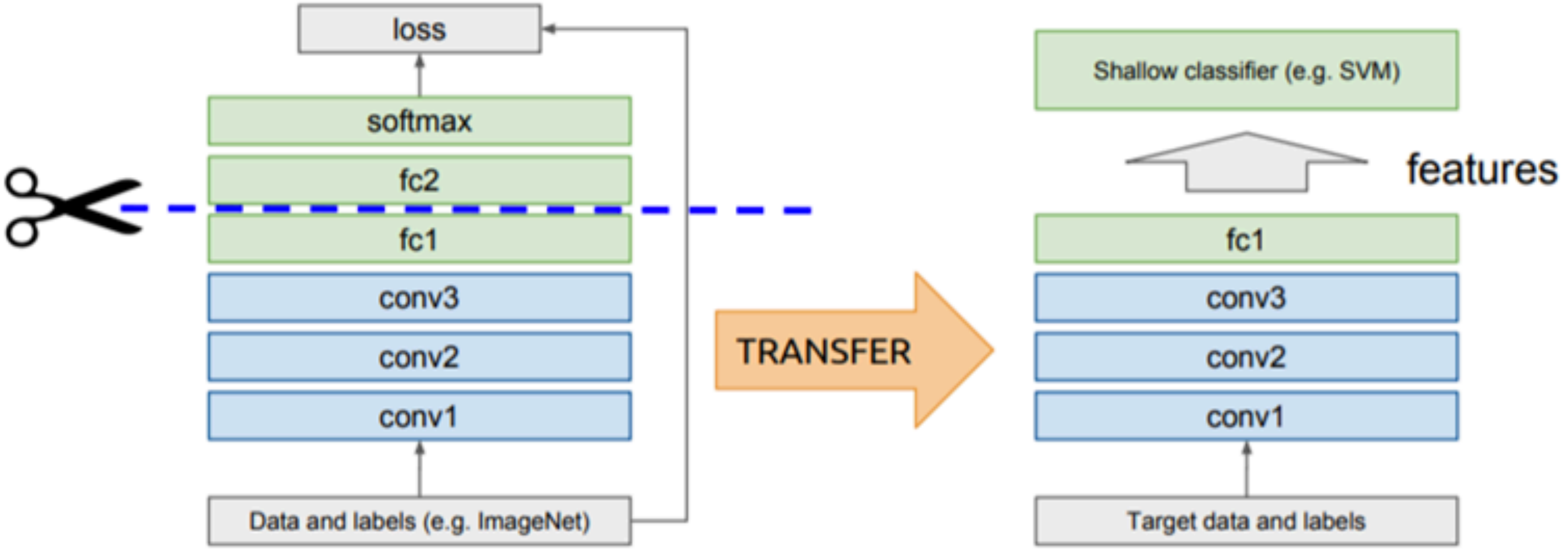
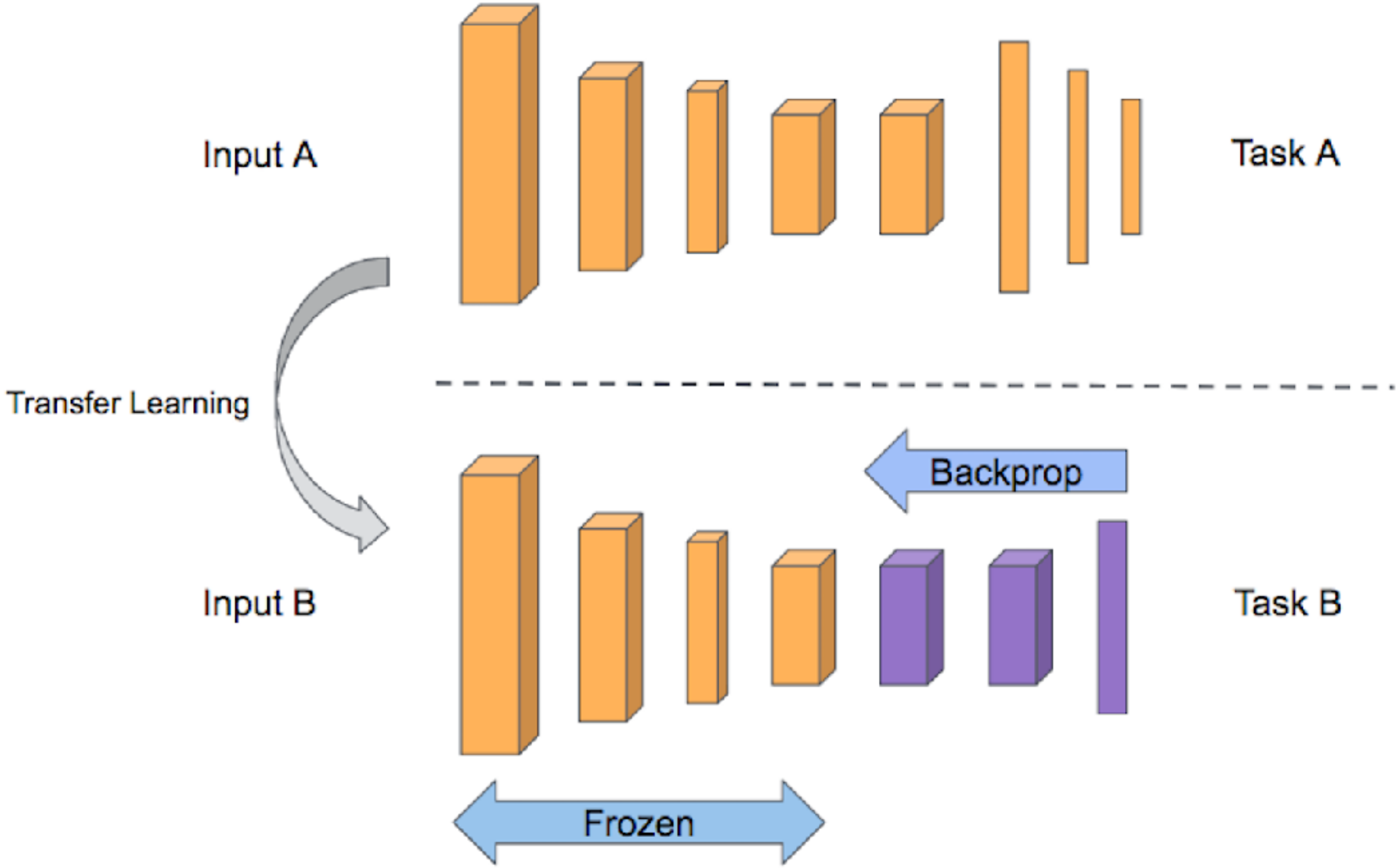


Image source: Dipanjan Sarkar/TowardsDataScience

Transfer Learning

Recall that the convolutional kernels are **learned** during the training process through backward propagation.
The trained kernels contains **knowledge** to detect patterns.
Why not let a new CNN to **inherit these knowledge**?



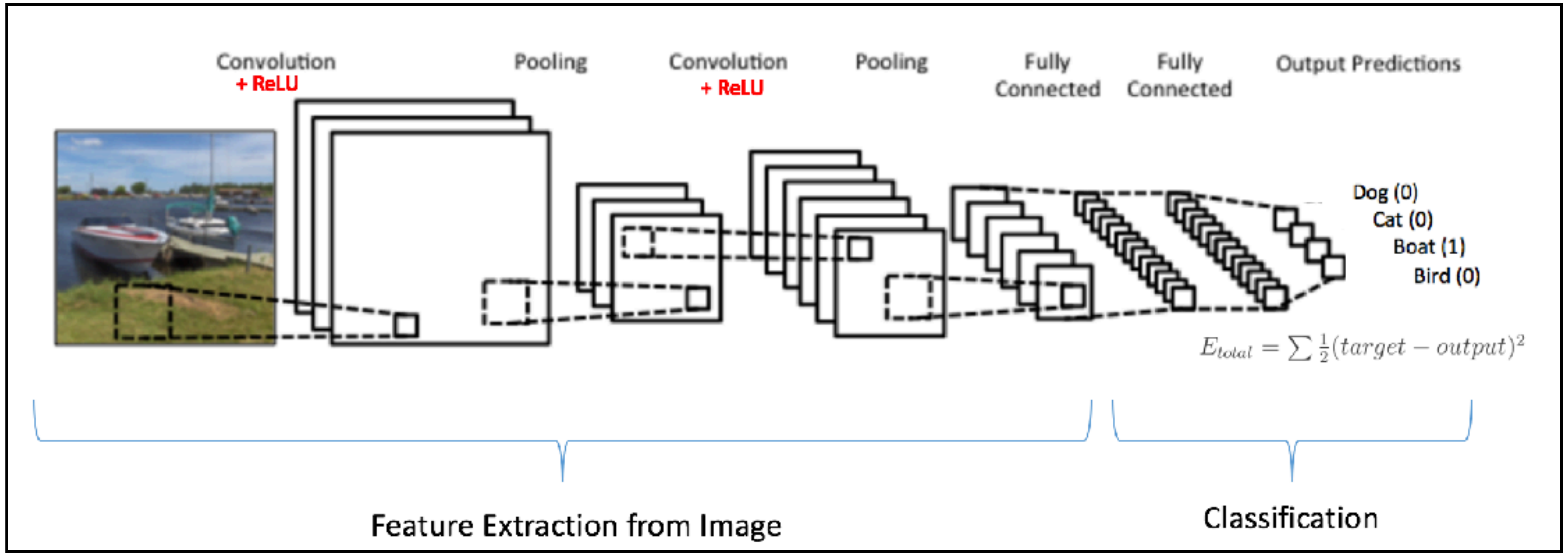

traffic
car 39%
car 25%
person 67%
car 27%
person
person 64%
person 56%
car 35%

Computer vision will be used by Mars Rover *Perseverance* to carry out automatic scientific experiments

Image credit: NASA

Take home messages

Convolutional neural networks (CNNs)
… are inspired by human eyes
… are based on the **filtering** technique in data processing
… implements filtering by applying the **convolution operation**
… encode the data **hierarchically** in an increasingly abstract way by layers
… are vital components of **computer vision**



A convolutional kernel
… is a **filter**
… is designed **automatically** during the training process by **backward propagation**
… is used for **feature extraction**
… is a **transferable knowledge**