Universität
Konstanz

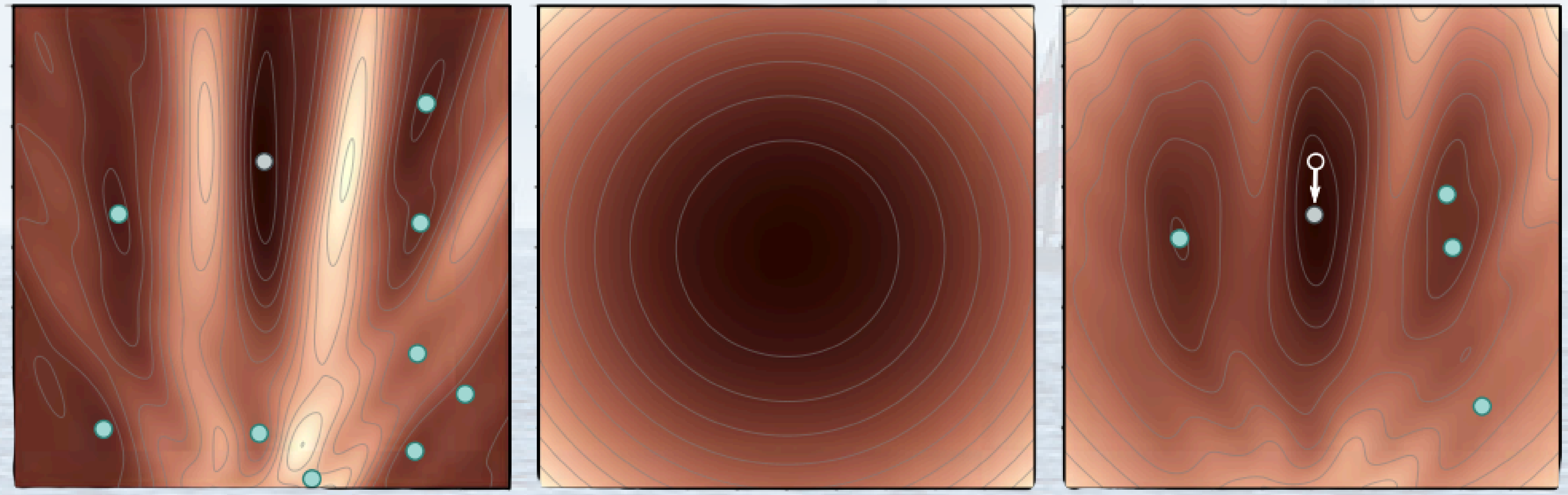
05 | Convolutional Neural Netwokrs

Giordano De Marzo
https://giordano-demarzo.github.io/

Deep Learning for the Social Sciences

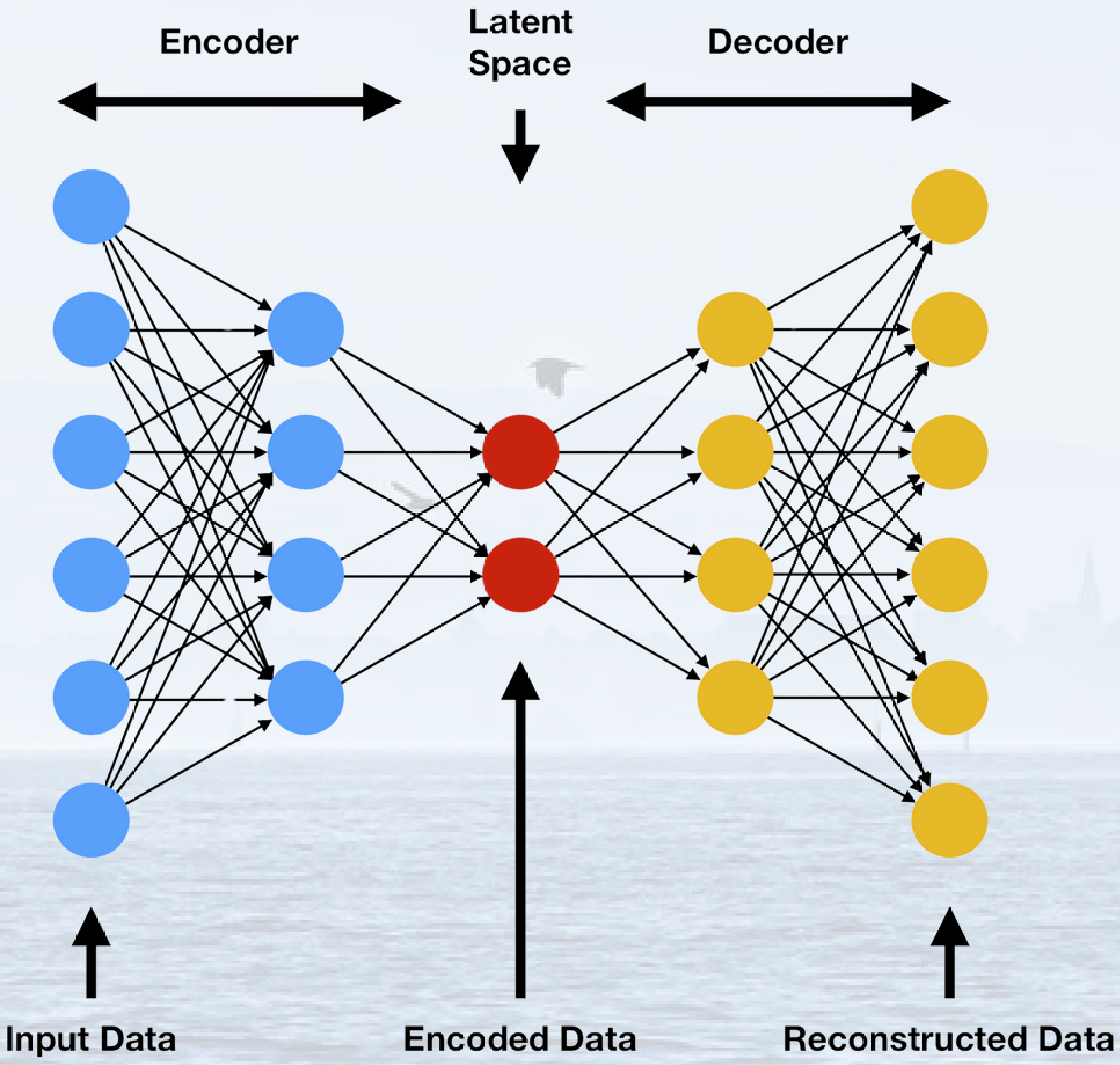
Regularization Techniques

Theoretically speaking a MLP with two hidden layers and many enough neurons can perform any arbitrary complex regression or classification task. In practice finding the best parameters achieving this is very hard. This is mainly due to the fact that the loss is a very irregular function with many local minima. Regularization techniques are used to make the loss more smooth and to improve the performances of DNN.



https://playground.tensorflow.org/

The Autoencoder



The Autoencoder is one of the most important MLP architectures for unsupervised learning. It is composed of three sections:

- **Encoder** Encodes the data into a latent representation
- **Latent Space** Space where the encoded data live
- **Decoder** Convert back the data from the latent space to the standard representation

The Autoencoder can be used for several tasks

- Dimensionality Reduction
- Anomaly Detection
- Denoising

Outline

1. Computer Vision
2. The Idea of Convolution
3. Convolutional Neural Networks

Universität
Konstanz

Computer Vision

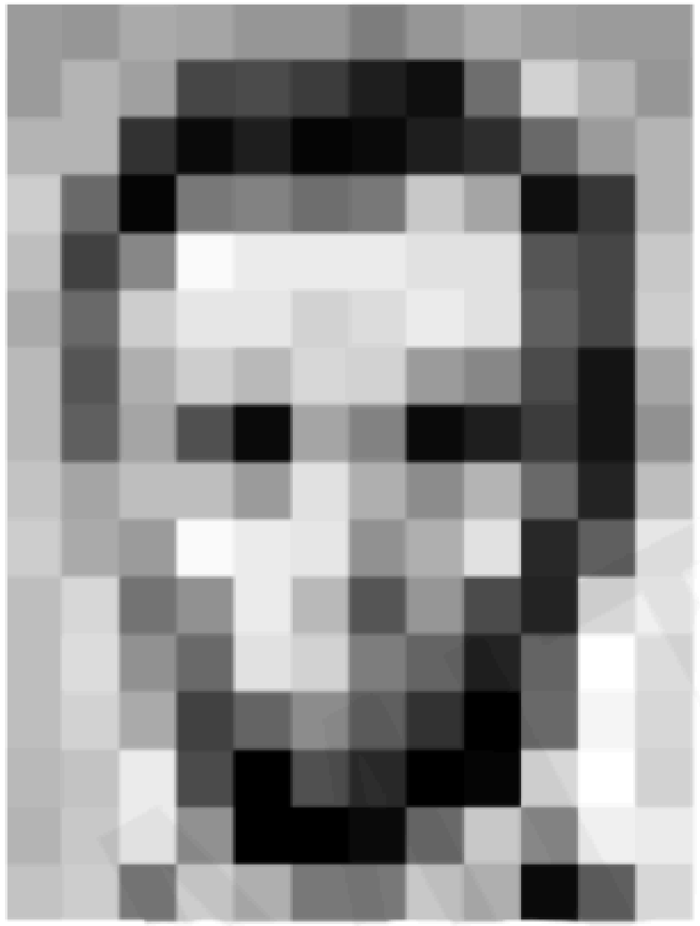
What is Computer Visions?

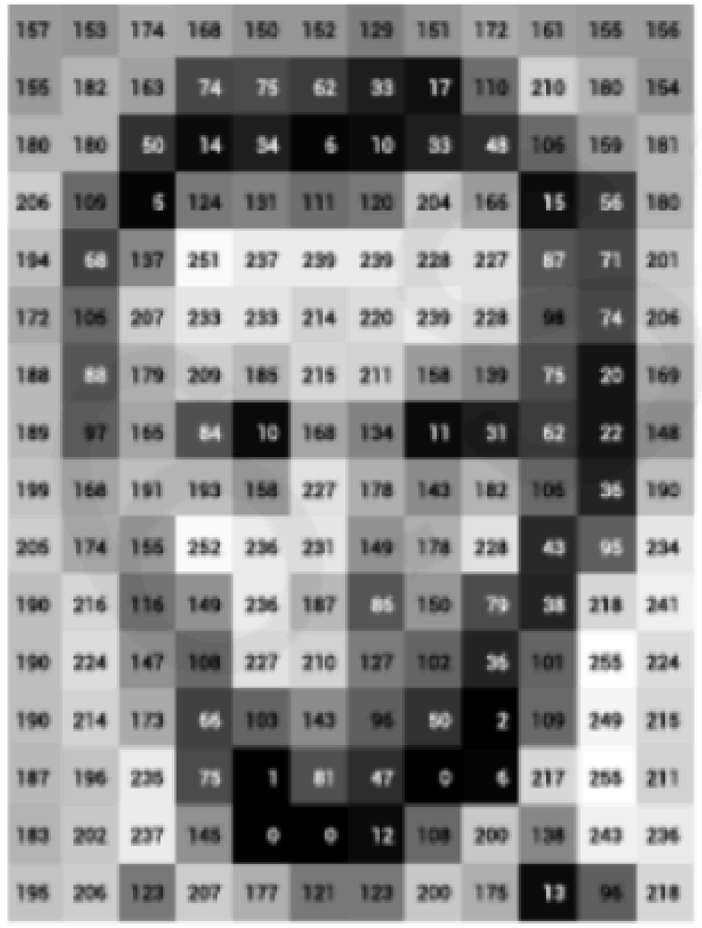
Computer vision is a field of artificial intelligence that focuses on enabling computers to interpret and understand digital images and videos

- Key tasks include image recognition, object detection, and segmentation
- Major challenges involve handling variations in lighting, image distortions, occlusions, noise, rotations etc
- Achieving reliable computer vision requires advanced deep learning models, significant computational power, and a lot of data

How Computers See Images

Images are stored on computers as matrices of pixels. Black and white pictures are represented as a single matrix storing numbers from 0 (black) to 255 (white)

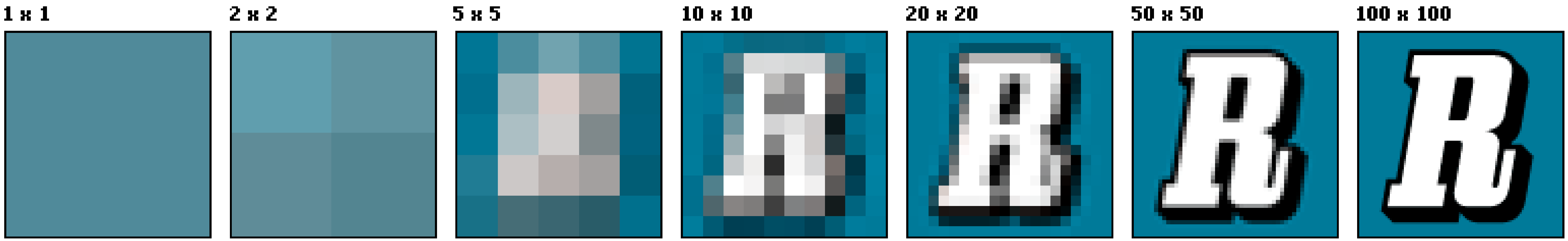




157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

Image Resolution

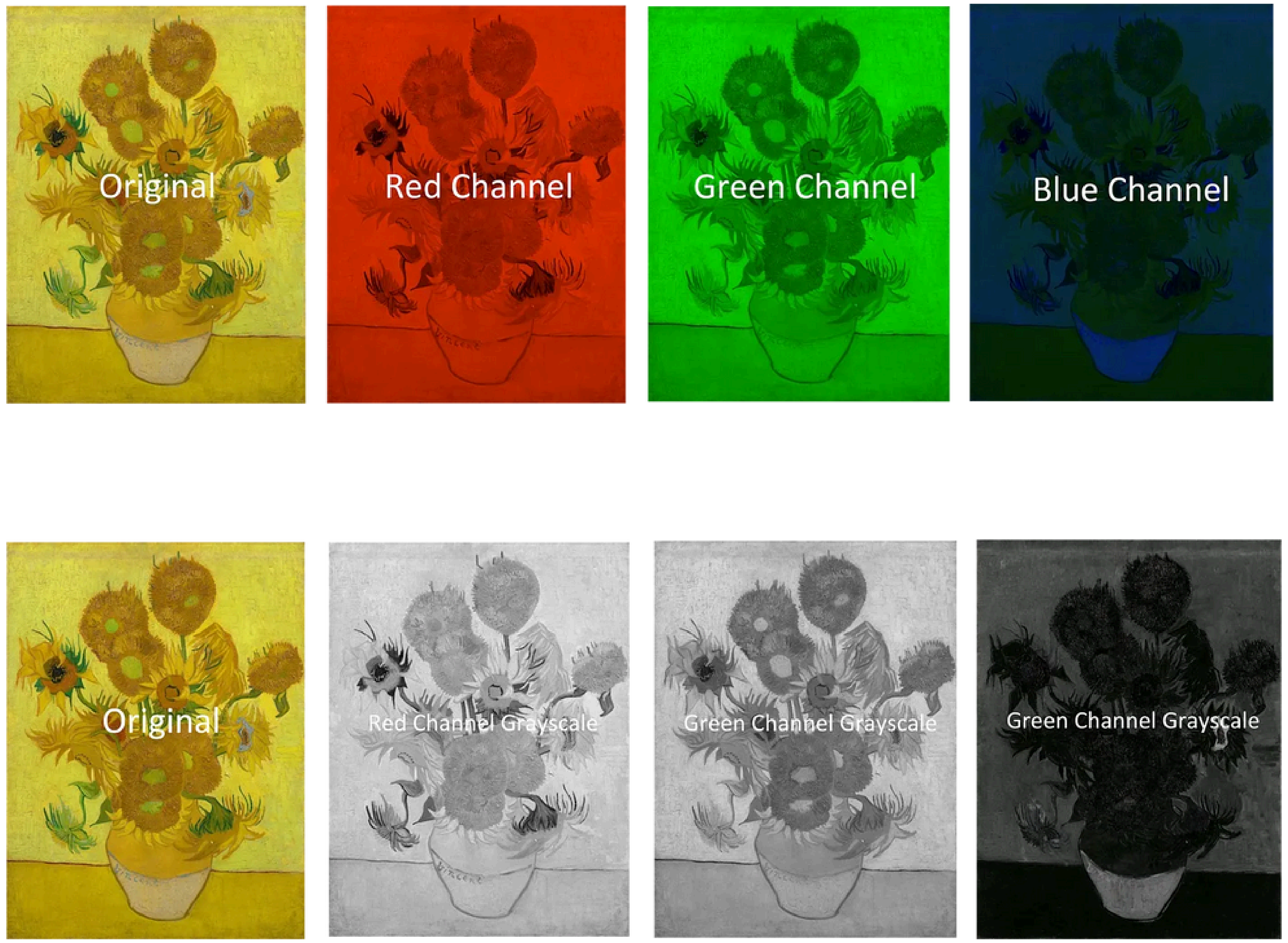
Image resolution refers to the amount of detail an image holds, typically described in terms of pixels. Pixels are the tiny, colored square components that make up an image. The resolution is often represented by the dimensions of the image in pixels, height H x width W. Each pixel corresponds to an entry of the matrix representing the image. Therefore, the larger the number of pixels or resolution, the higher will be the dimension of the input image.



Black and White vs Colors

Color images are represented digitally using a combination of color channels, typically red, green, and blue (RGB). Each channel stores intensity values for its respective color, and together, they create a full spectrum of colors when combined.

- **Black and White Images** Represented as one matrix, where each entry corresponds to a pixel
- **Color Images** Represented as three matrices, one for each channel (color). The elements of the matrices corresponds to pixels and the values in the matrices give the intensity of the three colors in each location of the image



Example of Computer Vision Applications

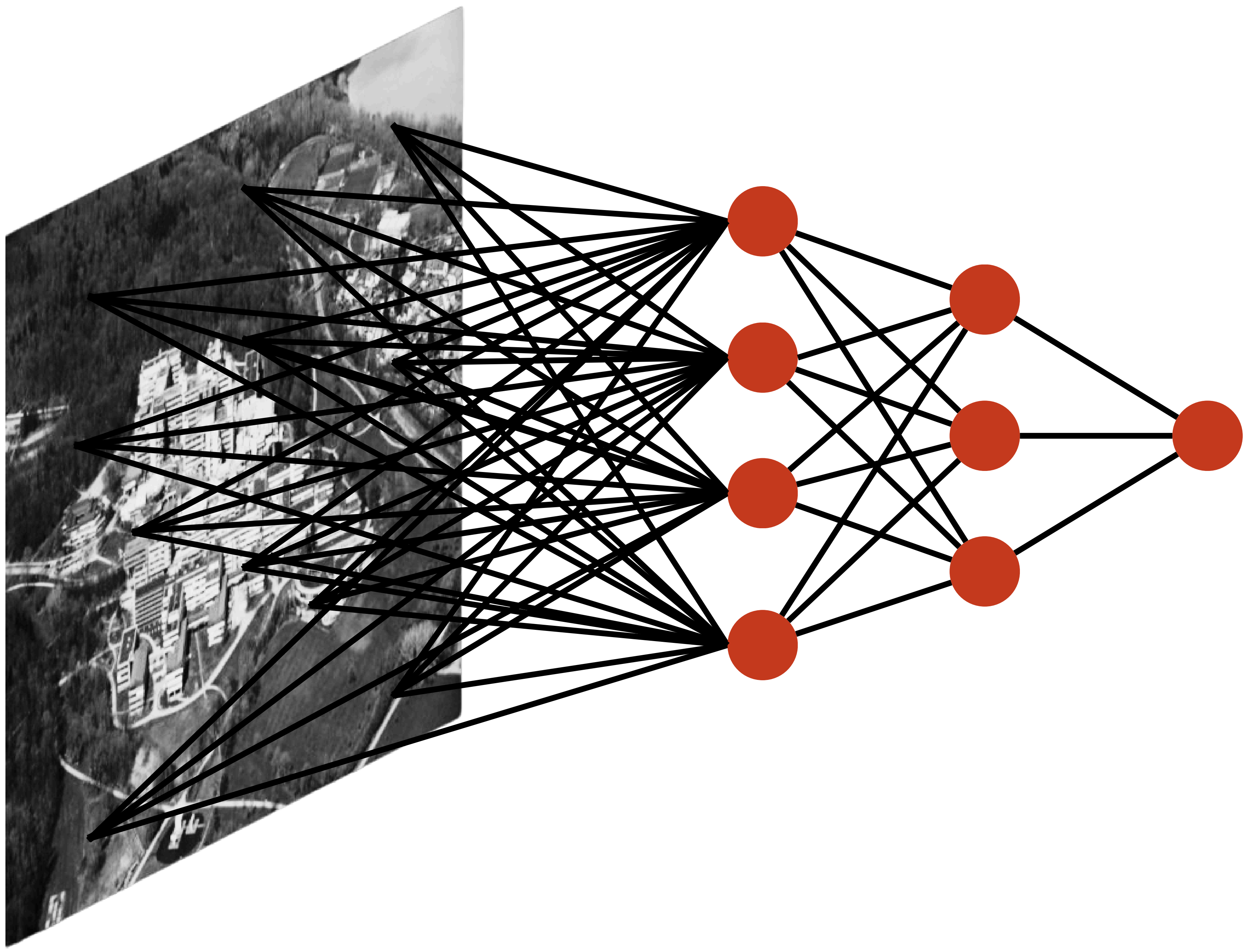
Computer vision technology has a wide range of applications

- **Autonomous Vehicles:** Computer vision systems enable cars to perceive their surroundings, detect pedestrians, recognize traffic signs, and make navigation decisions, contributing to safer autonomous driving.
- **Facial Recognition:** Used widely in security and surveillance, mobile phone unlocking, and photo tagging on social platforms, facial recognition technology relies on computer vision to identify individuals based on their facial features.
- **Healthcare:** In medical imaging, computer vision helps in diagnosing diseases, analyzing X-rays, MRIs, and CT scans, and supporting surgeries by providing precise visual guidance.
- **Agriculture:** Advanced computer vision algorithms are employed to monitor crop health, manage pests, and automate harvesting processes, significantly increasing efficiency and productivity.

Universität Konstanz

Convolution and Filters

Using MLPs on Images

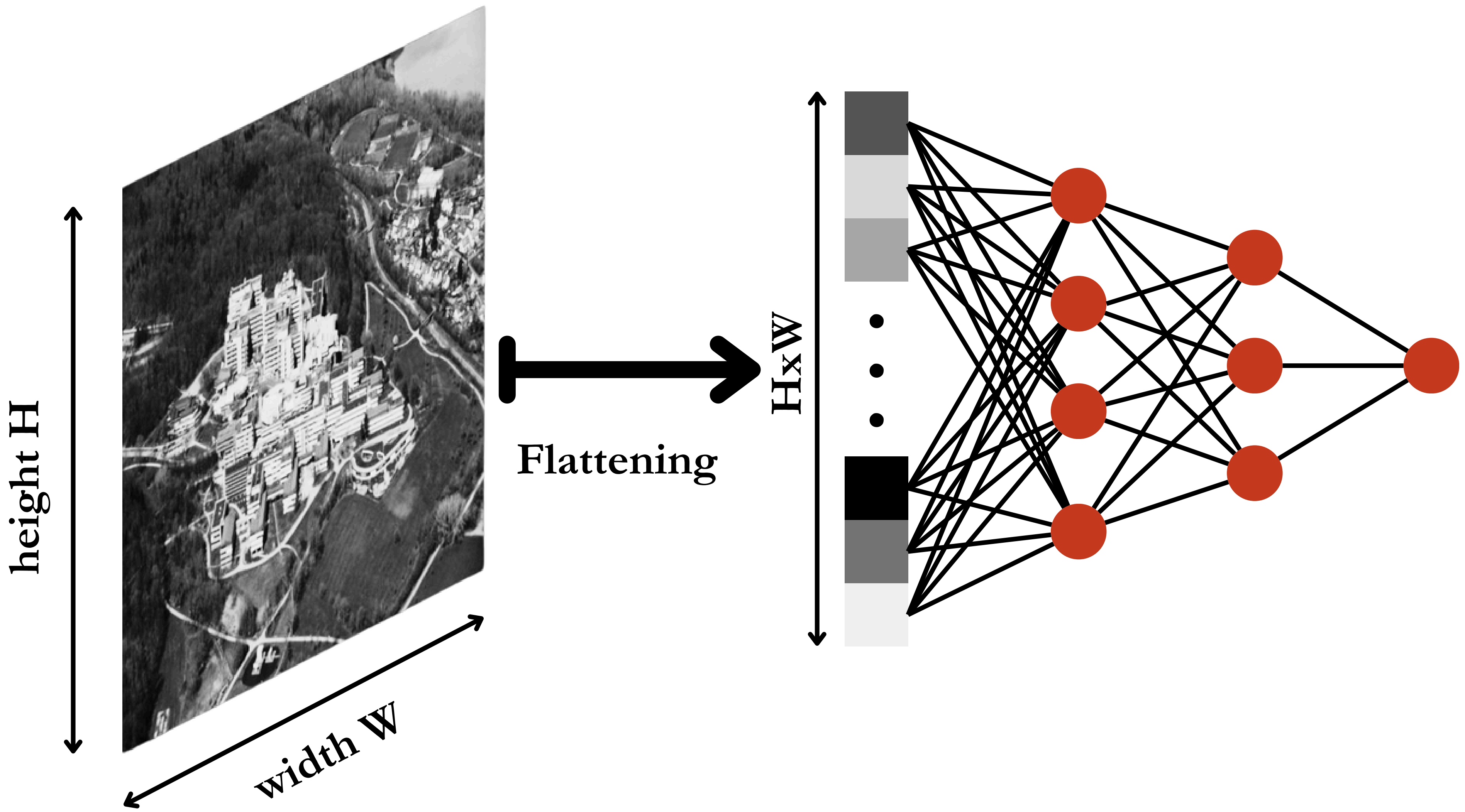


We know that a sufficiently complex MLP can perform any task, including computer vision ones

- each pixel of the input matrix corresponds to an entry of the input layer
- each neuron in the first hidden layer is connected to all pixels on the input image
- if we have P pixels and H hidden neurons, we have PxH connections
- we can then proceed with the standard MLP architecture

Flattening Layers

In order to feed images into a MLP we use a flattening layer. It is a layer with no parameters, it only transforms the image from a matrix to a linear vector.



Limits of MLPs on Images

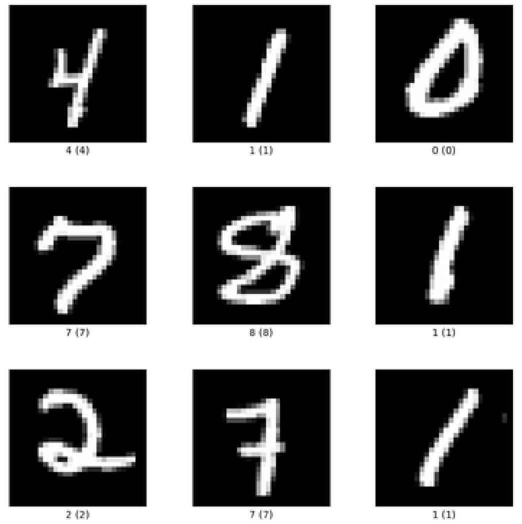
In theory a sufficiently complex MLP can handle images, in practice this is almost impossible for two main reasons:

- there are too many parameters, especially in high resolution images, coming from the first hidden layer
- the flattening layer destroys all the spatial structure in the image

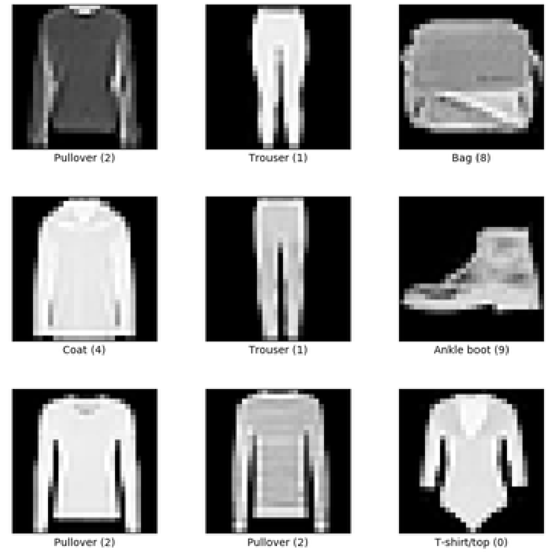
For this reason MLPs on images achieve good performances only on very simple tasks:

- MNIST 97-98% accuracy
- Fashion MNIST 85-89% accuracy
- CIFAR-10 40-50% accuracy
- CIFAR-100 20-30% accuracy

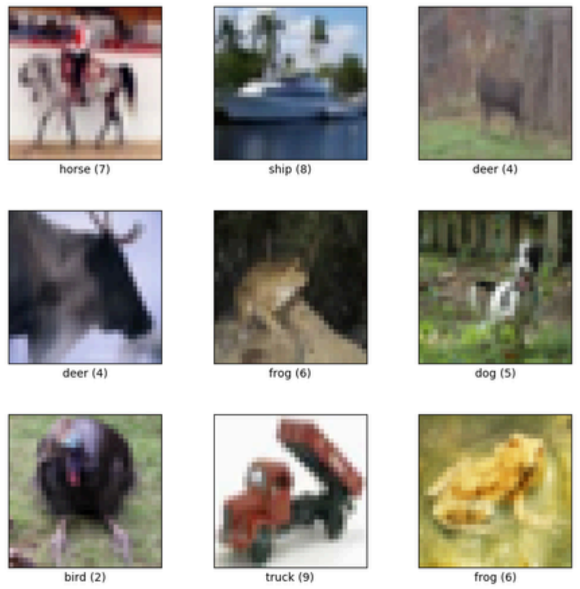
MNIST



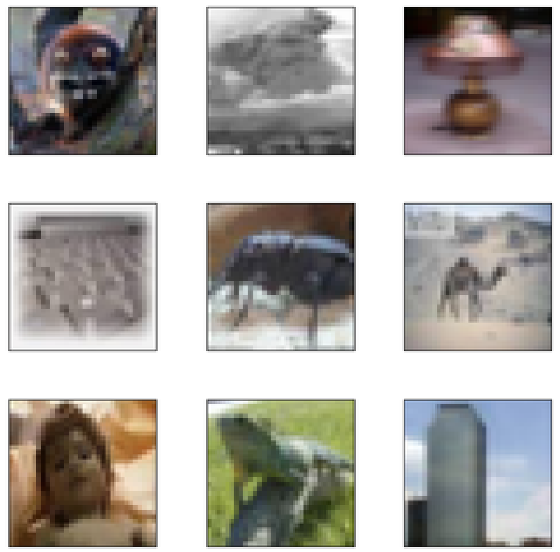
Fashion MNIST



CIFAR-10



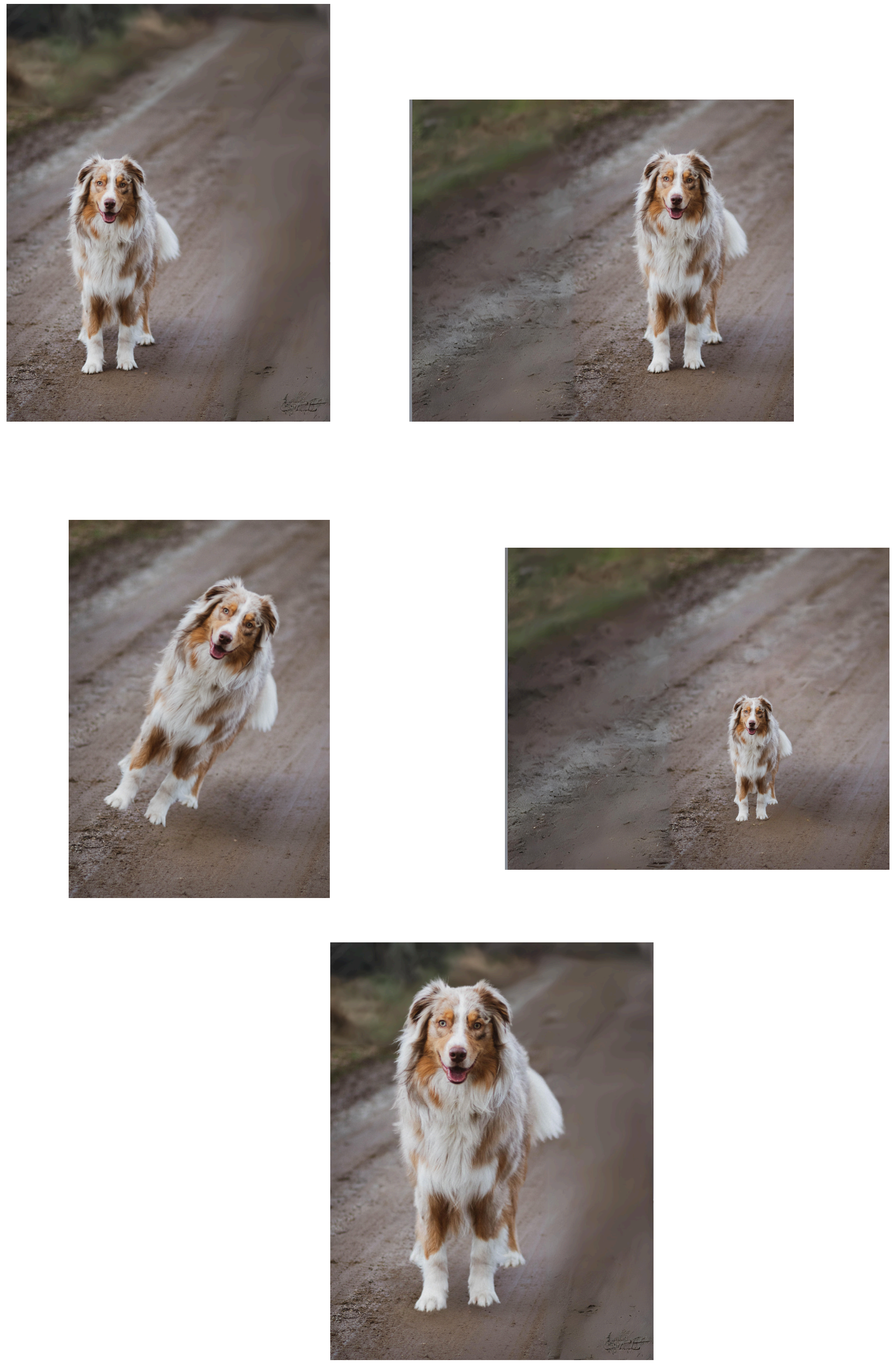
CIFAR-100



Properties of Images

The MLP considers images as flattened monodimensional input vectors, neglecting all the relevant properties of image. Images:

- have a structure and spatial correlations
- are rotational invariant: in a classification task the specific orientation of the object does not matter
- are translational invatiant: in a classification task the specific placement of the object in the image does not matter
- can be deformed or shrunk, but this should not affect our neural network



Convolution

1	0	1
0	1	0
1	0	1

Filter/Kernel

1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
0	0	1	1	0
0	1	1	0	0

Input

Convolution is an operation that involves two matrices:

- a small matrix called filter or kernel
- a larger input matrix

The convolution consists in

- sliding the filter over the input image
- performing element-wise multiplication
- summing the results of the multiplications

Convolution

Filter/Kernel

1	0	1
0	1	0
1	0	1

Input

1x1	1x0	1x1	0	0
0x0	1x1	1x0	1	0
0x1	0x0	1x1	1	1
0	0	1	1	0
0	1	1	0	0

Output

4		

Universität Konstanz

Convolution

1	0	1
0	1	0
1	0	1

Filter/Kernel

1	1x1	1x0	0x1	0
0	1x0	1x1	1x0	0
0	0x1	1x0	1x1	1
0	0	1	1	0
0	1	1	0	0

Input

4	3	

Output

Convolution

1	0	1
0	1	0
1	0	1

Filter/Kernel

1	1	1x1	0x0	0x1
0	1	1x0	1x1	0x0
0	0	1x1	1x0	1x1
0	0	1	1	0
0	1	1	0	0

Input

4	3	4

Output

Convolution

1	0	1
0	1	0
1	0	1

Filter/Kernel

1	1	1	0	0
0x1	1x0	1x1	1	0
0x0	0x1	1x0	1	1
0x1	0x0	1x1	1	0
0	1	1	0	0

Input

4	3	4
2		

Output

Convolution

Filter/Kernel

1	0	1
0	1	0
1	0	1

Input

1	1	1	0	0
0	1	1	1	0
0	0	1x1	1x0	1x1
0	0	1x0	1x1	0x0
0	1	1x1	0x0	0x1

Output

4	3	4
2	4	3
2	3	4

Examples of Filters

Original

Sharpen

Edge Detect

Strong Edge Detect

Stride

In the example the filter was moving pixel by pixel. We call stride the number of pixels the filter moves.

1	0	1
0	1	0
1	0	1

1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
0	0	1	1	0
0	1	1	0	0

Stride=1

1	0	1
0	1	0
1	0	1

1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
0	0	1	1	0
0	1	1	0	0

Stride=1

1	0	1
0	1	0
1	0	1

1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
0	0	1	1	0
0	1	1	0	0

Stride=2

1	0	1
0	1	0
1	0	1

1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
0	0	1	1	0
0	1	1	0	0

Stride=2

1	0	1
0	1	0
1	0	1

1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
0	0	1	1	0
0	1	1	0	0

Padding

Padding consists in adding a border to the image. In this way we can modify the output size of the convolution and we don’t loose information around the borders and corners:

- typically the border is made of zeros (zero-padding)
- the thickness is typically one, but larger paddings can be used

0	0	0	0	0	0	0
0	1	1	1	0	0	0
0	0	1	1	1	0	0
0	0	0	1	1	1	0
0	0	0	1	1	0	0
0	0	1	1	0	0	0
0	0	0	0	0	0	0

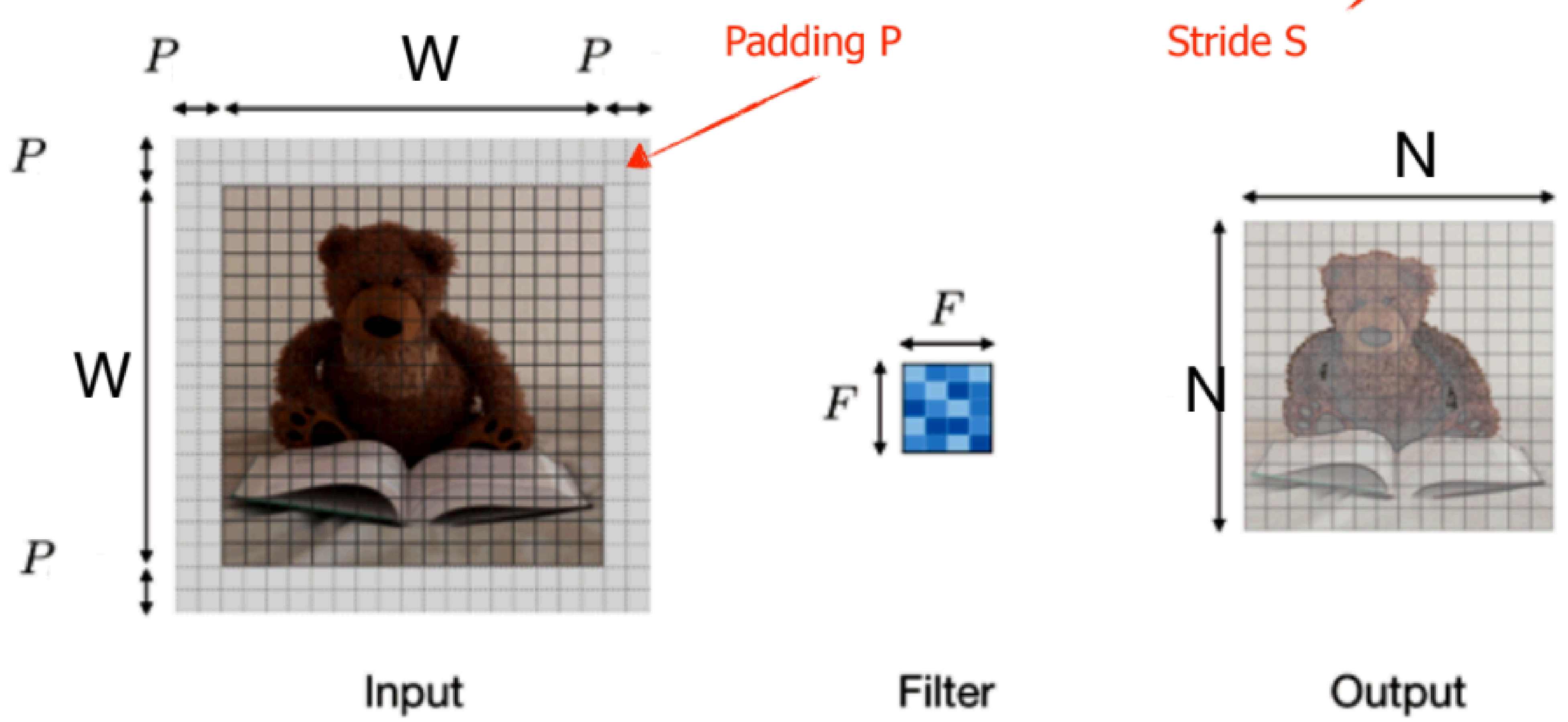
Output Dimension

The output dimension depends on the input size, the size of the filter, the padding and the stride

https://poloclub.github.io/cnn-explainer/

Size of the output after applying the filter:

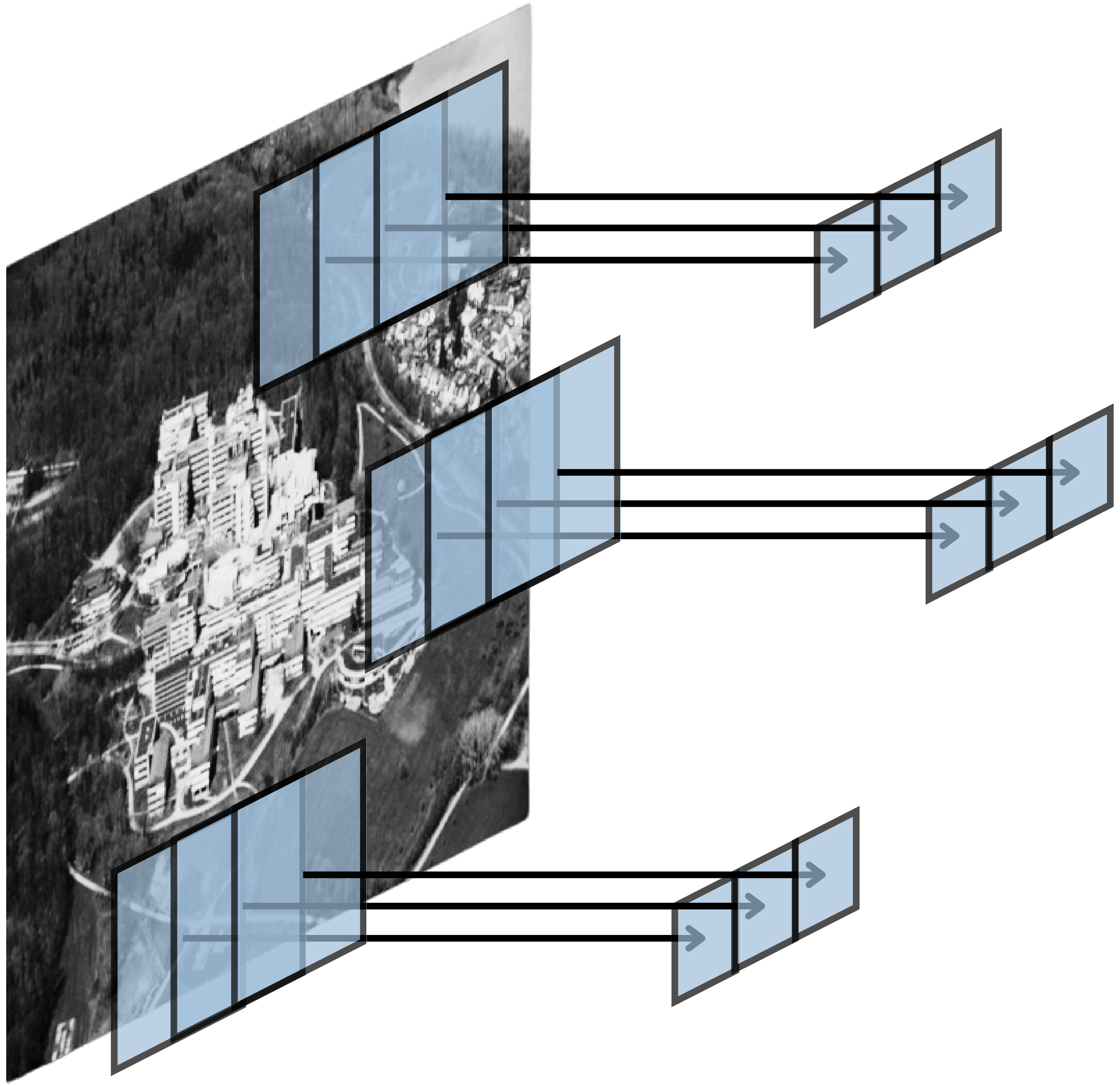
$$\boxed{N = \frac{W - F + 2P}{S} + 1}$$



Universität Konstanz

Convolutional Neural Networks

Learning Filters



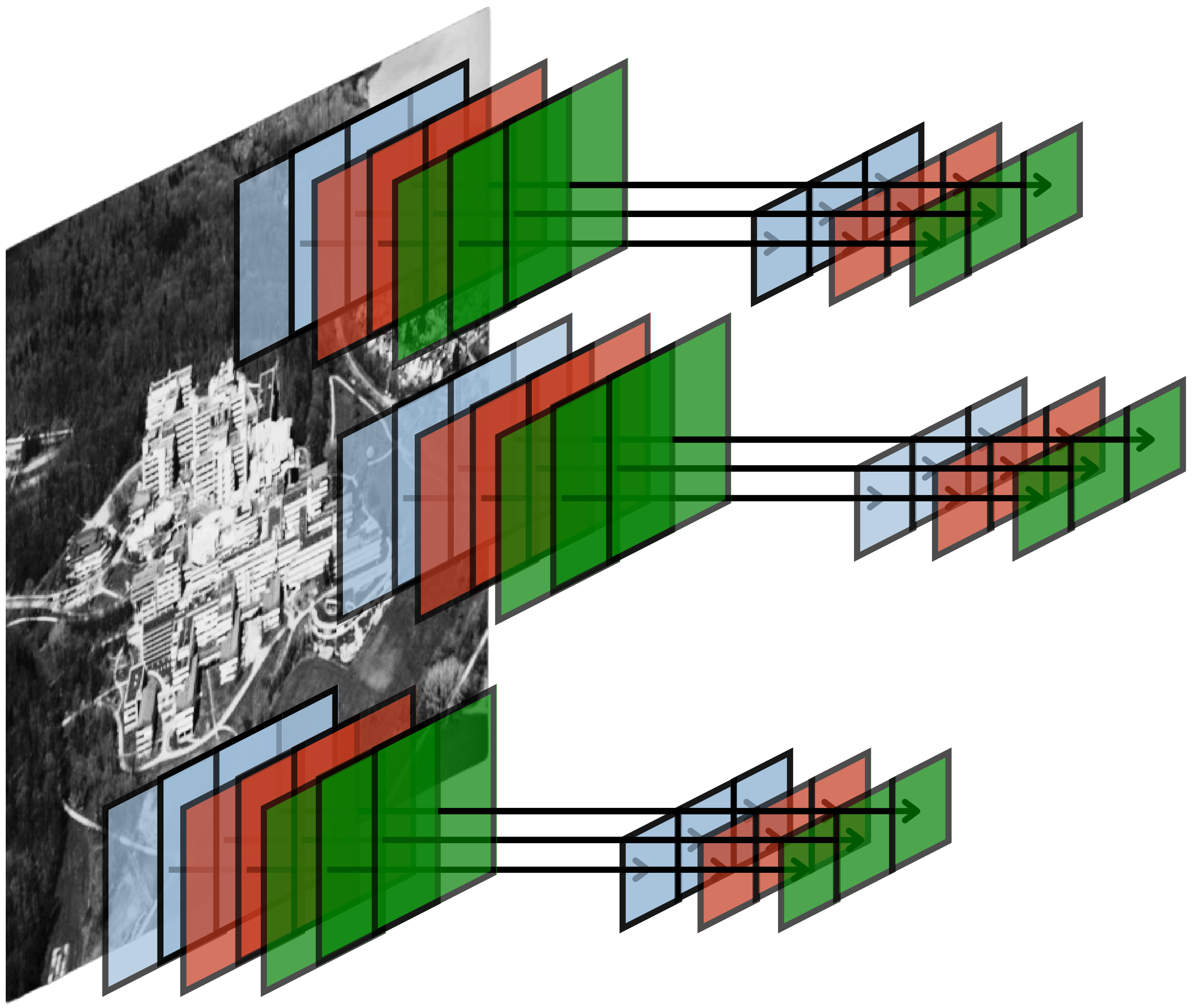
Convolutional Neural Networks are based on a very simple idea:

The parameters we want to learn are the numbers inside the filter

This is a very powerful approach:

- we reuse parameters because the same filter is applied to all the locations of the images
- the filter captures local geometrical features, such as edges, independently of where they are placed in the image
- the output is another (smaller) matrix that can be further processed to extract more complex feattures

Using Many Filters

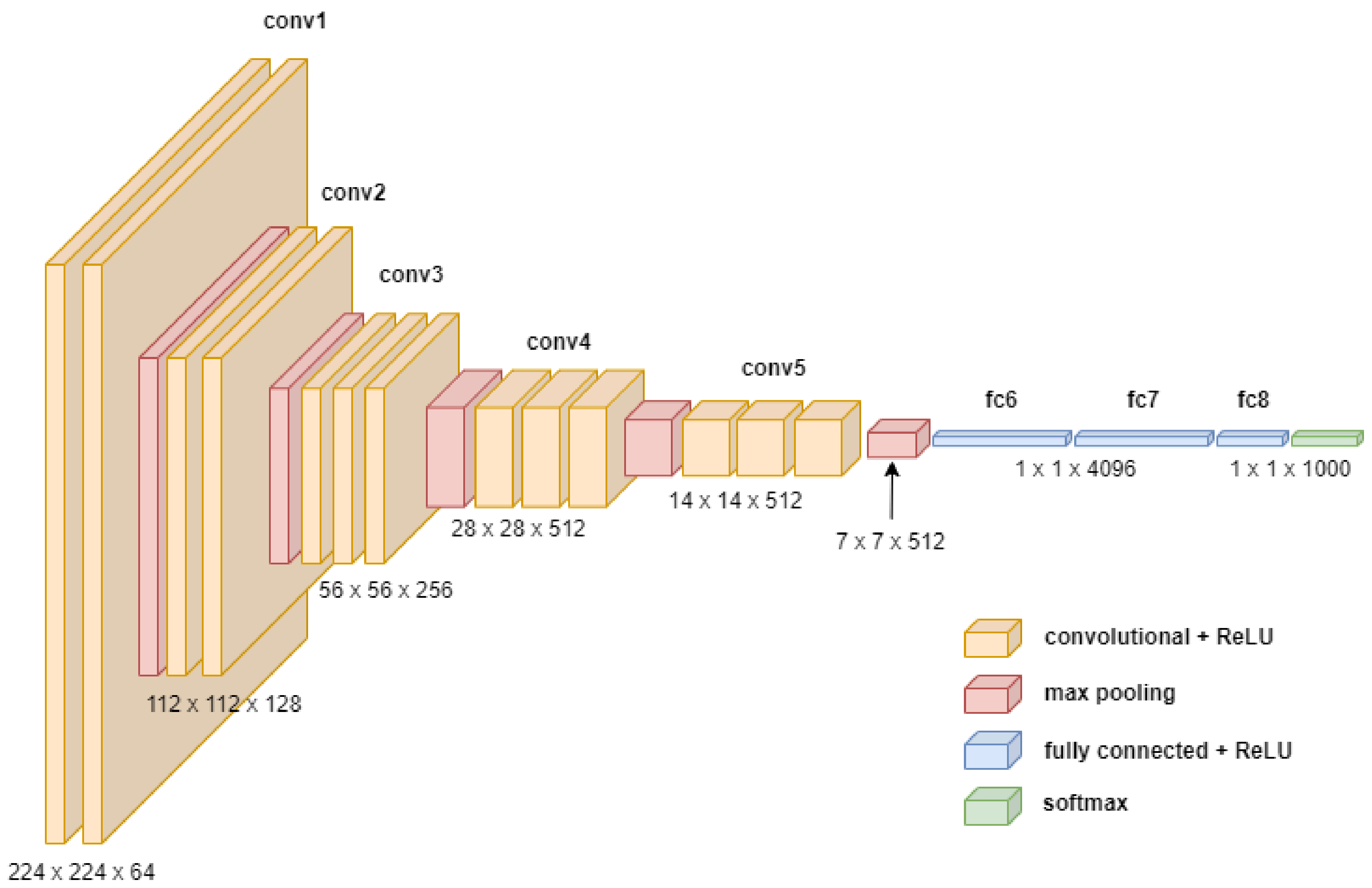


Just one filter is not enough for learning complex features:

- we use many different filters
- during the learning process the numbers in each filter are updated
- each filter will specialize in a different task:
    - detecting edges
    - detecting vertical lines
    - detecting horizontal lines

Even if we use many filters, we don't have many parameters, because we are reusing them in different locations of the image.

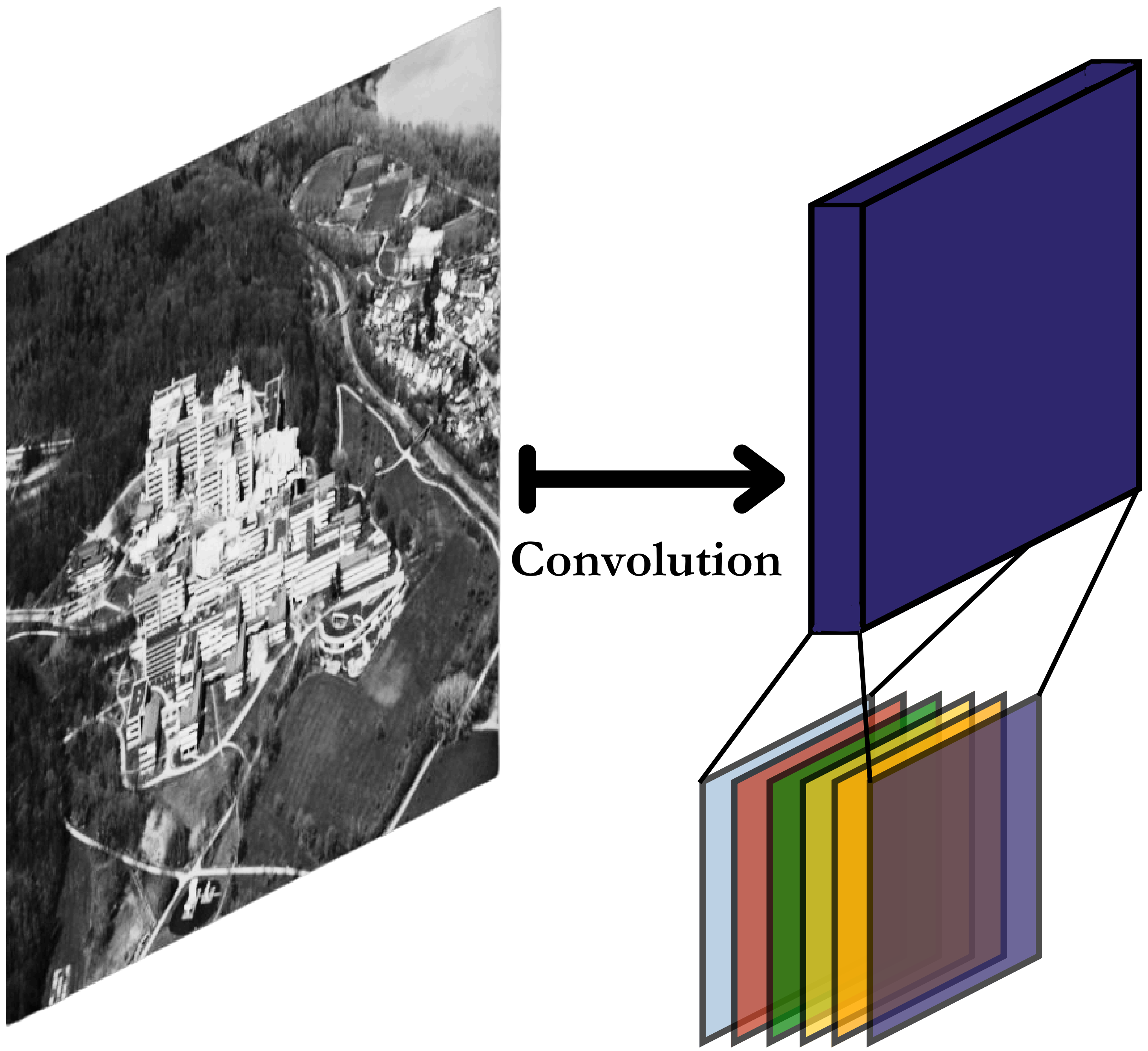
Schema of a Convolutional Neural Network



A Convolutional Neural Network (CNN) is composed of three types of layers:

- convolutional layers, responsible for extracting visual features from the image
- pooling layers, that reduce the dimension, increasing robustness and diminishing the number of parameters
- fully connected layers, that use the features extracted by the convolutional layers to perform regression or classification

Convolutional Layer

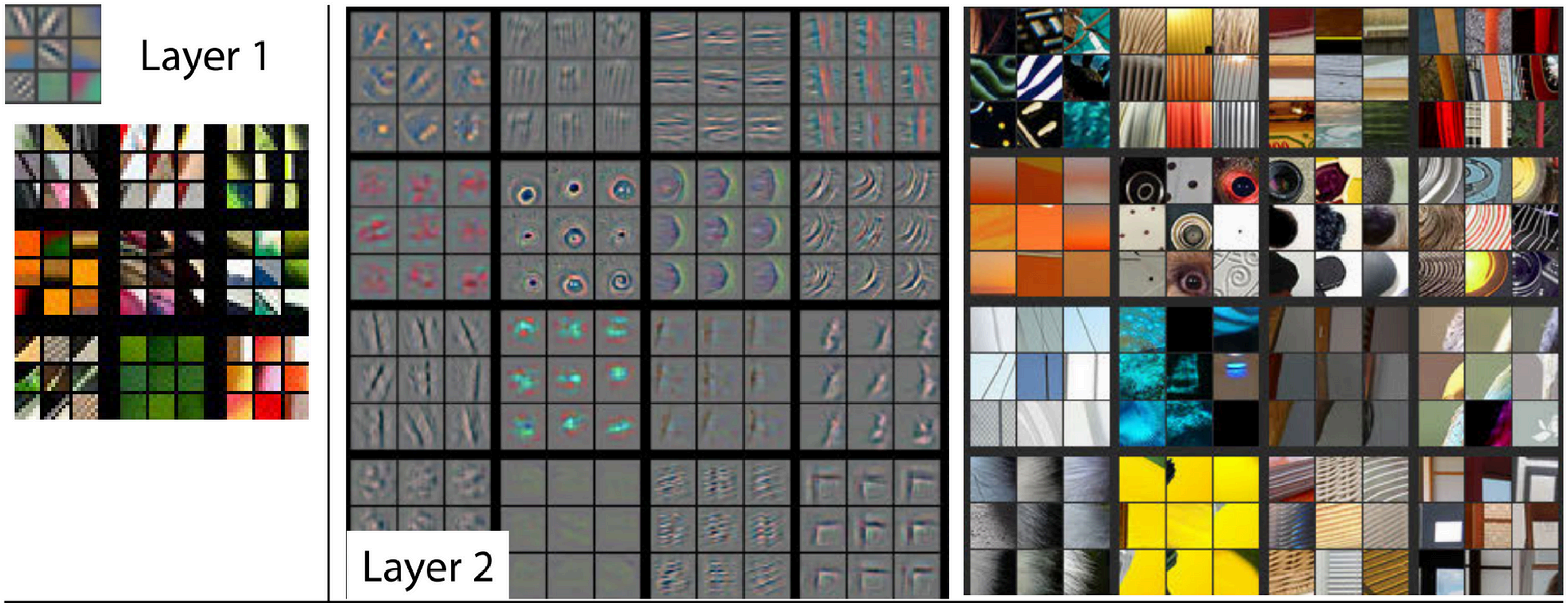


A convolutional layer consists of F different filters:

- it takes in input an image of dimension WxH
- it applies each filter to the input image
- the result of each filter is passed through an activation function (ReLU)
- the output consists of F stacked images , one for each filter
- each output image captures different features of the original image
- the hyperparameters are: number of filters, filters size, stride and padding

What are the Filters Learning?

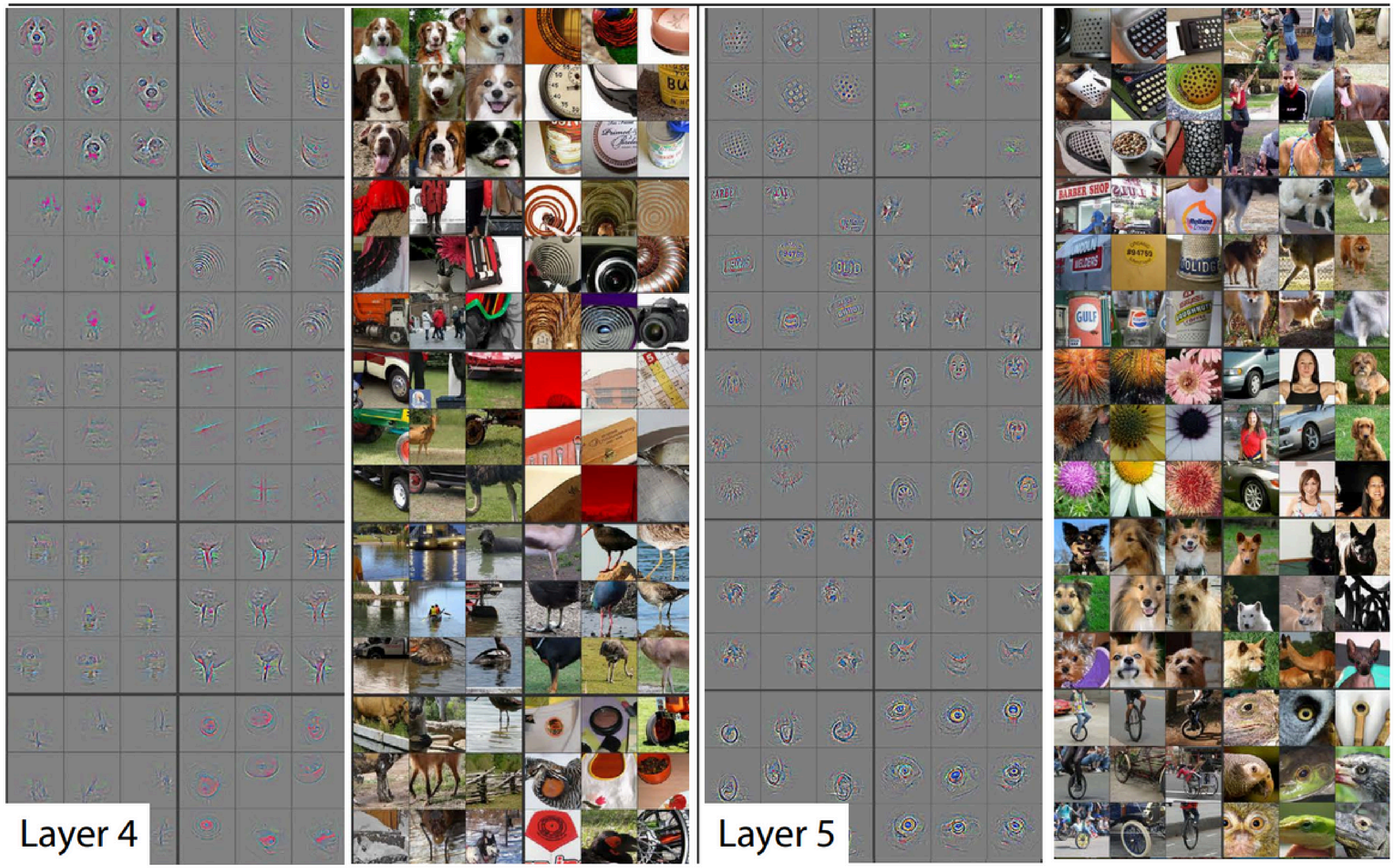
Initial layers capture lines and simple shapes



Universität
Konstanz

What are the Filters Learning?

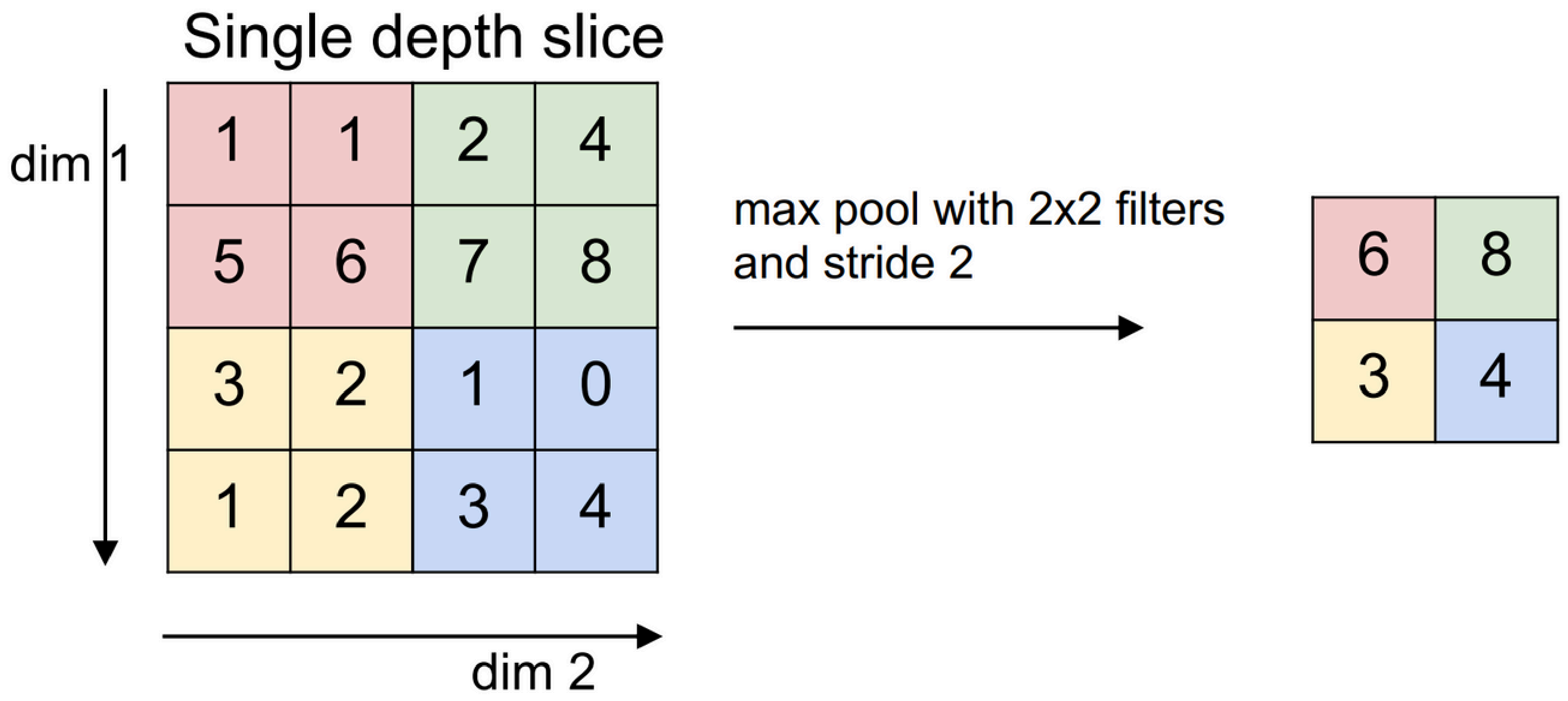
Deeper layers extract more complex features such as faces.



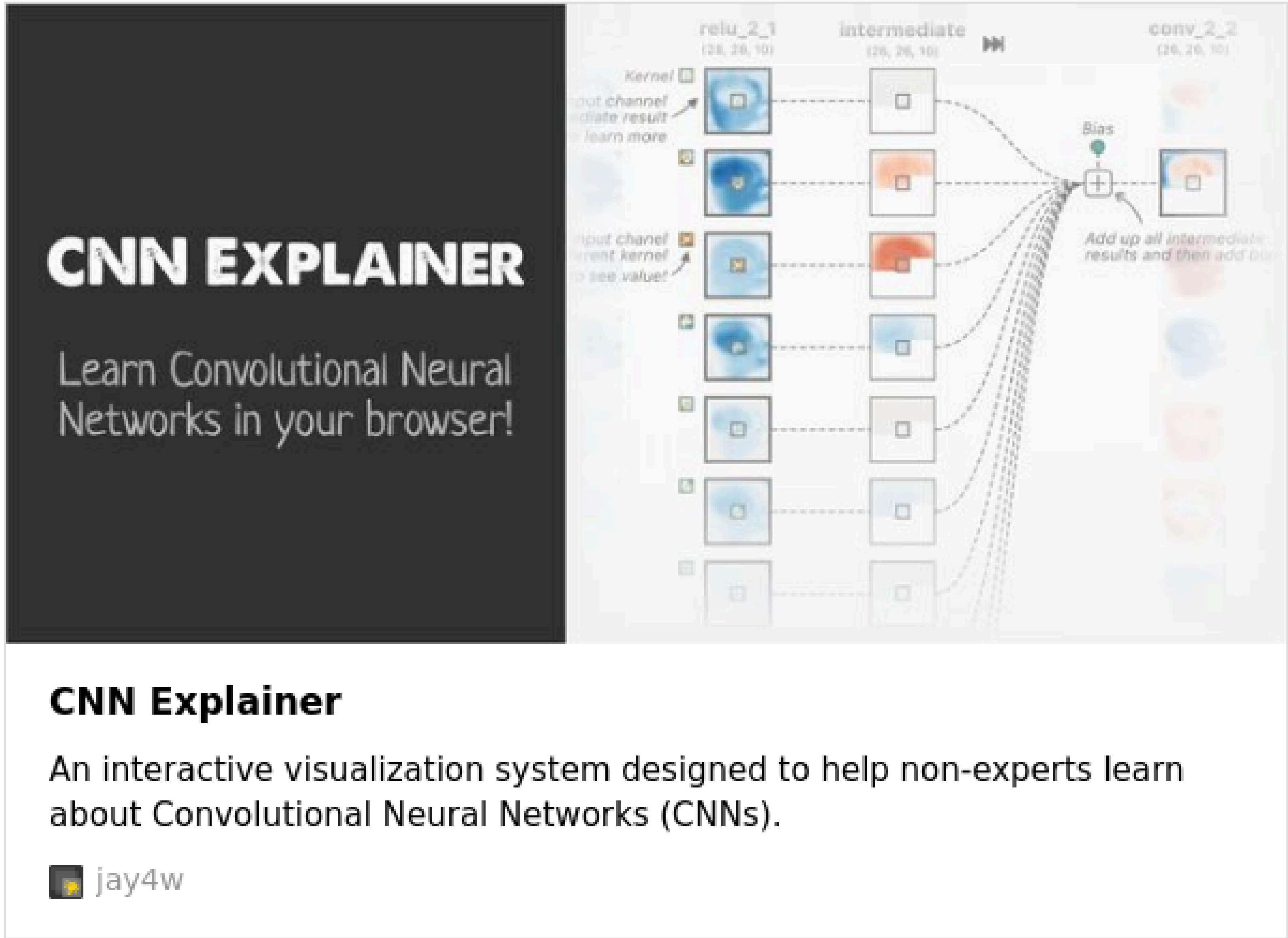
Pooling Layers

Pooling Layers reduce the dimension of the output of convolutional layers.

- The output of each filter is divided in small blocks of PxP pixels
- For each block, only one value is given in output
- Different possible approaches:
    - max pooling
    - average pooling



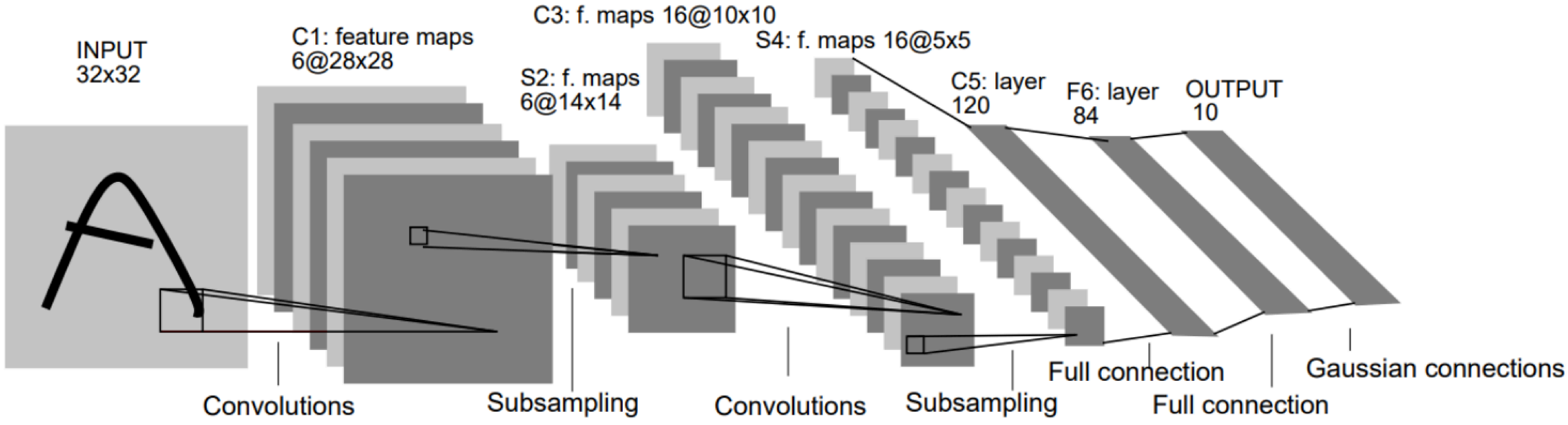
CNN Examples



https://cs.stanford.edu/people/karpathy/convnetjs/demo/mnist.html

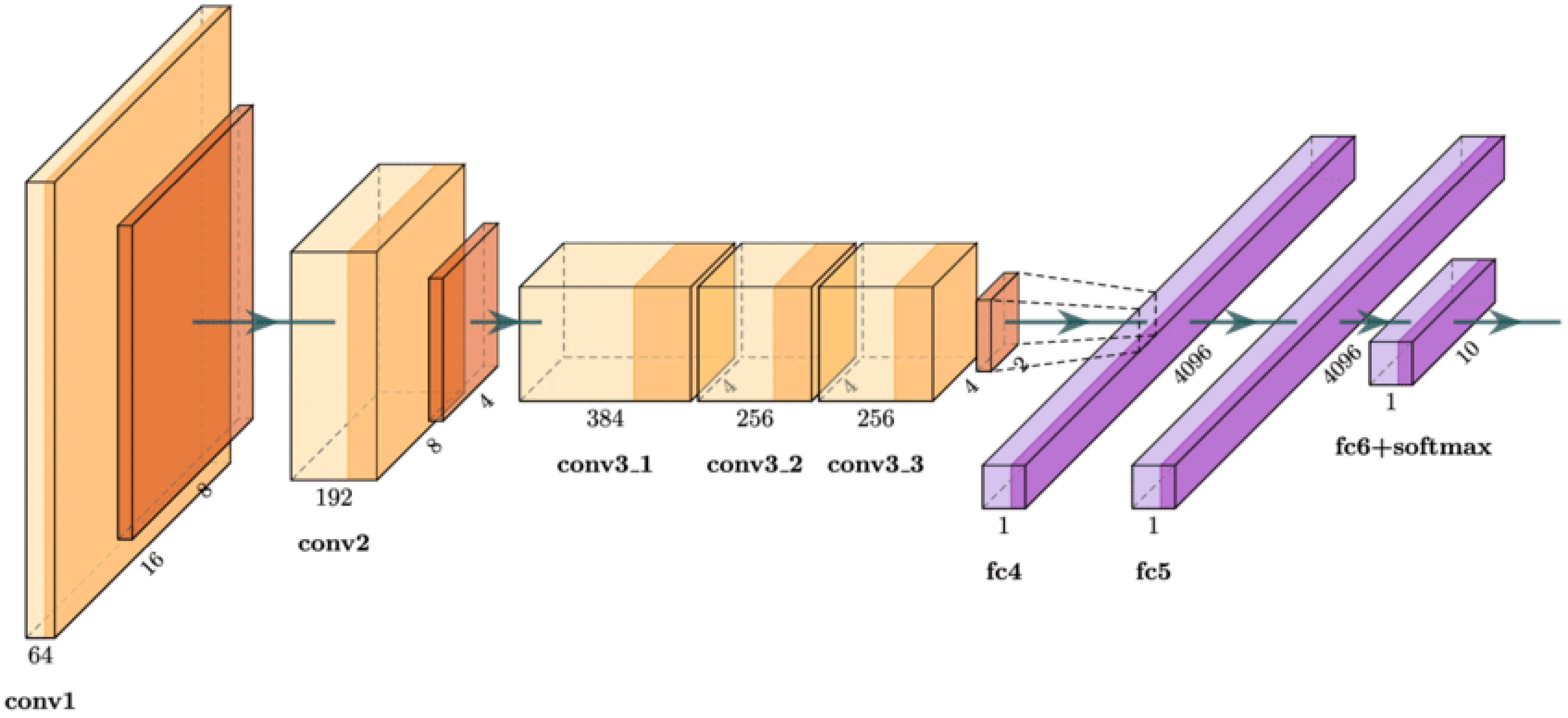
LeNet-5

LeNet-5 (1998) is one of the first examples of CNN. It consists of 2 convolutional layers, two pooling layers and 3 fully connected layers. It was created with the aim of performing digits recognition.



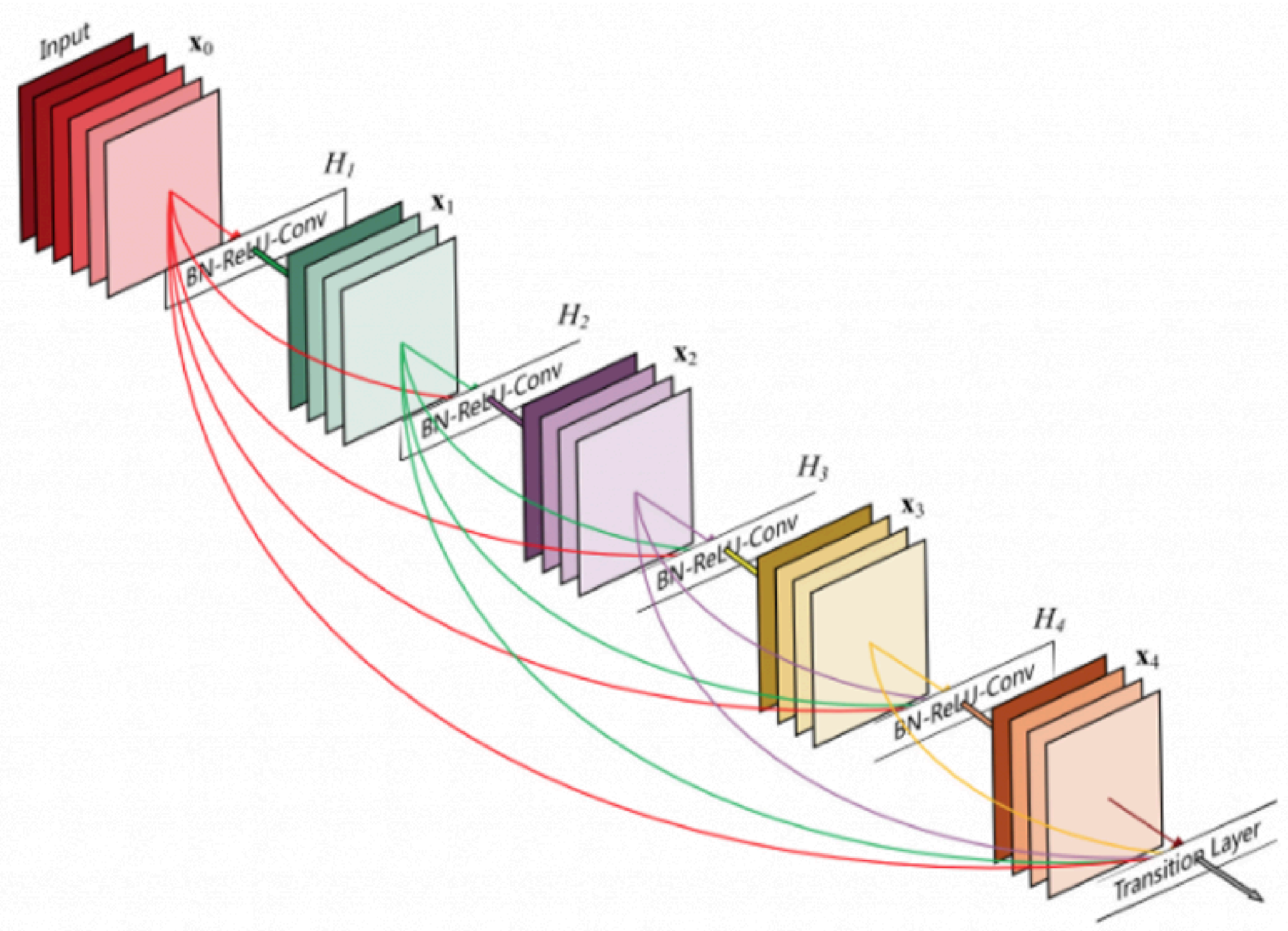
AlexNet

AlexNet (2012) has been the first CNN to achieve state of the art performances on image classification tasks. Even few convolutional layers outperforms much larger MLP based architectures.



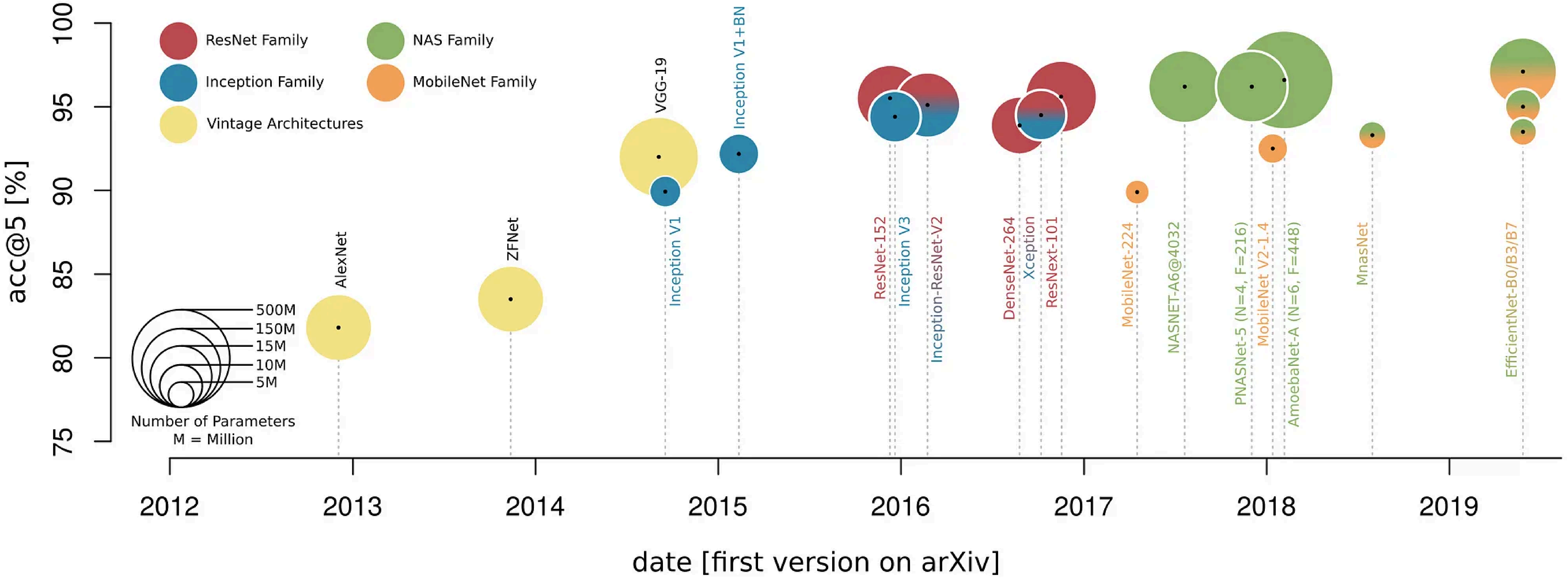
More Advanced CNNs

As convolutional neural networks (CNNs) grow deeper, they often face challenges like vanishing gradients and degradation problems, where adding more layers leads to higher training error. To address these issues, ResNet (Residual Network) introduces "residual connections". This is achieved by adding shortcut connections that bypass one or more layers.



Universität Konstanz

SOTA CNN Models



Summary

Computer Vision

- Computer vision is a field of artificial intelligence that focuses on enabling computers to interpret and understand digital images and videos

Convolution and Filters

- Using MLPs on images in not feasible
- Images are rotational and translational invariant, can be deformed or shrunk, light may change
- Convolution is a matrix operation that allows to extract features such as edges from images

Convolutional Neural Networks

- Instead of learning weights we learn the filters
- Deeper filters learn more complex features

Next Lectures and Events

Tomorrow Morning CDM Colloquium (08/05 - Room D301 10:30-11:30)

Max Pellert from Barcelona Supercomputing Center will give a talk on using LLMs for surveys. The title is "Synthetic Surveys for Population Insights"

Tomorrow Afternoon Coding Session (08/05)

In the coding session we will code two convolutional neural networks, one for classifying clothes, the other for classifying animals

Next Week

We will introduce Graph Neural networks, a very powerful tool to analyze graph data. Raphaela Keßler will connect from Munich to talk about her master thesis. She focused on using Graph Neural Networks to detect misinformation on Telegram.