

Oliver Dürr

Short course on deep learning
Block 2



Southern Taiwan University
of Science and Technology

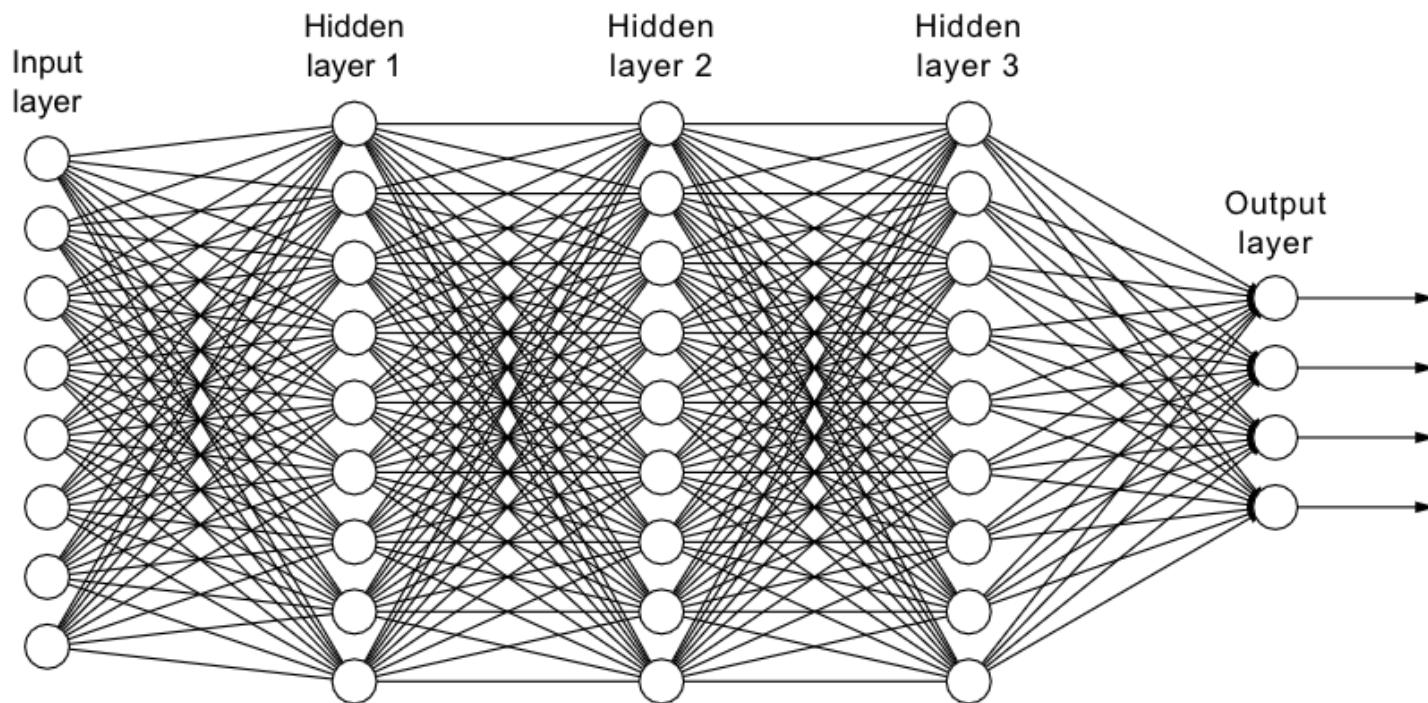
This course is a short version of dl_course_2022 I created with
Beate Sick.
Hochschule Konstanz

09.11.22

Topics of today

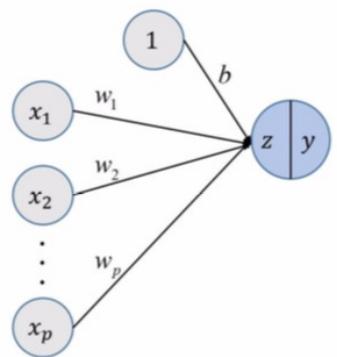
- A second look on fully connected Neural Networks (fcNN)
- Convolutional Neural Networks (CNN) for images
 - Motivation for switching from fcNN to CNNs
 - Introduction of convolution
 - ReLu and Maxpooling Layer
 - Biological inspiration of CNNs
 - Building CNNs

Architecture of a fully connected NN

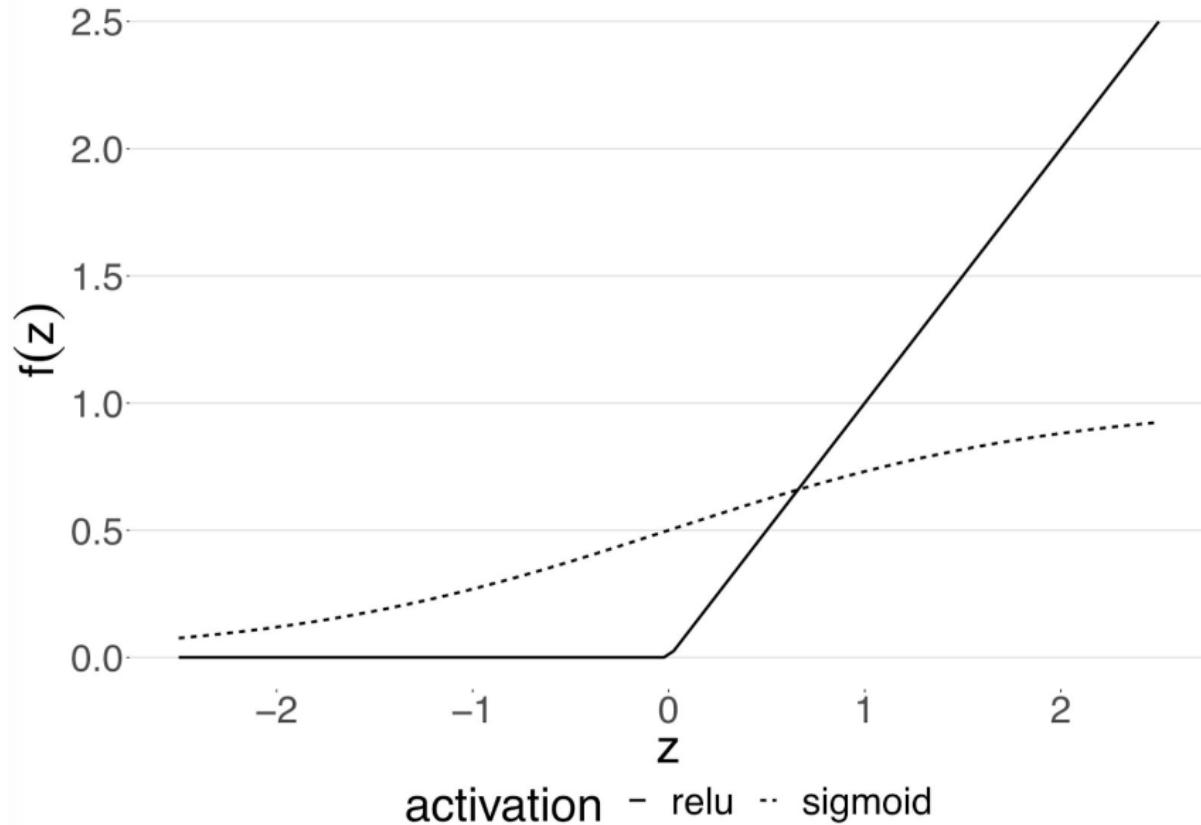


Each neuron in a fcNN gets as input a weighted sum of all neuron activation from one layer below. Different neurons in the same layer have different weights in this weighted sum, which are learned during training.

Common non-linear activation function

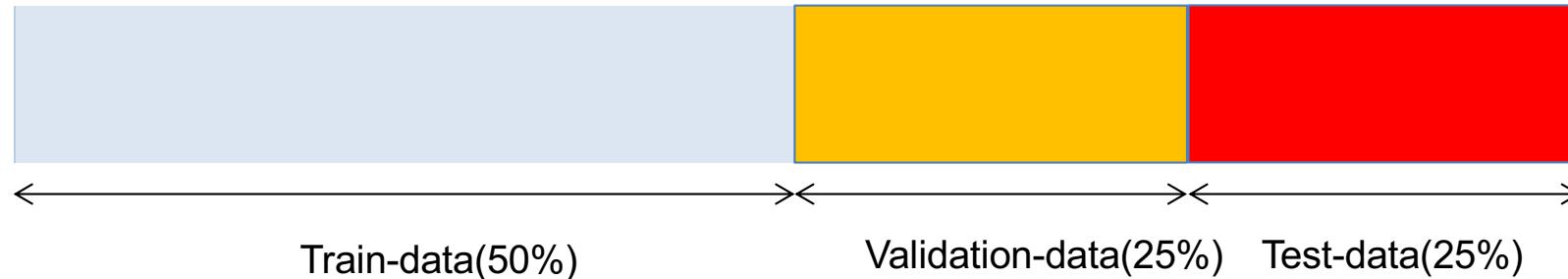


$$y = f(z) = f(b + \sum x_i \cdot w_i)$$



ReLU leads to faster training.,.

Best practice: Split in Train, Validation, and Test Set



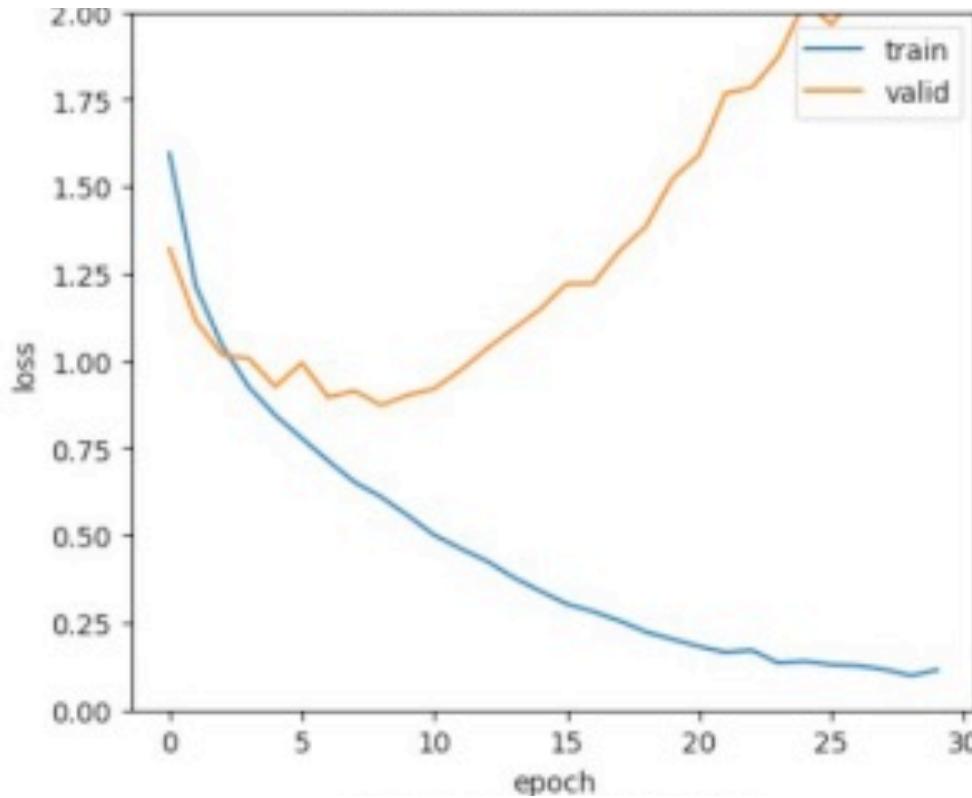
Best practice: Lock an extra **test data set** away, and use it only at the very end, to evaluate the chosen model, that performed best on your validation set.

Reason: **When trying many models, you probably overfit on the validation set.**

Determine performance metrics, such as MSE, to evaluate the predictions **on new validation or test data**

What can loss curves tell us?

Very common check: Plot loss in train and validation data vs epoch of training.

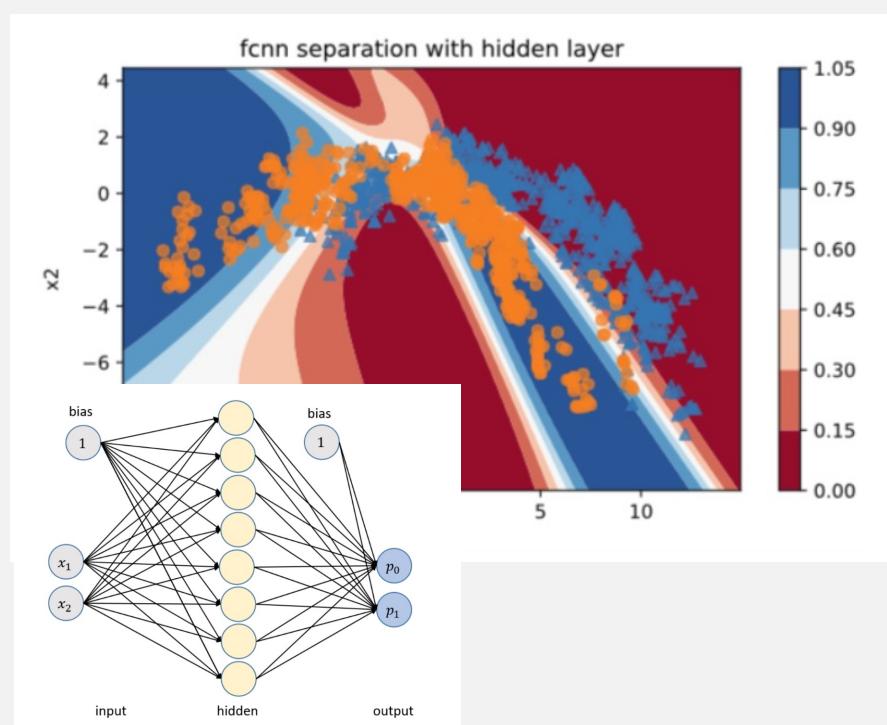
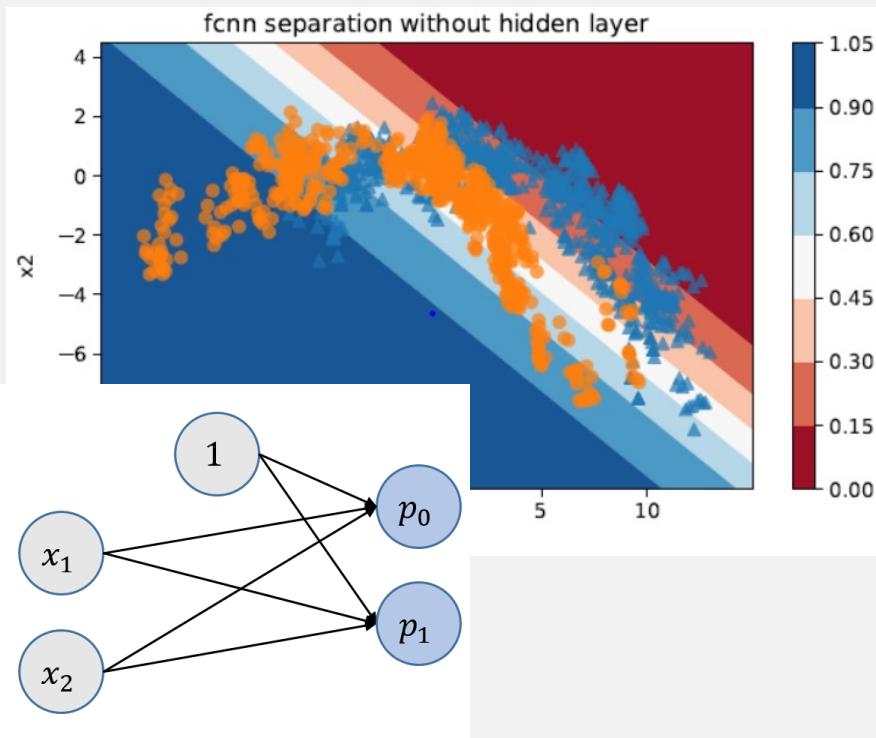
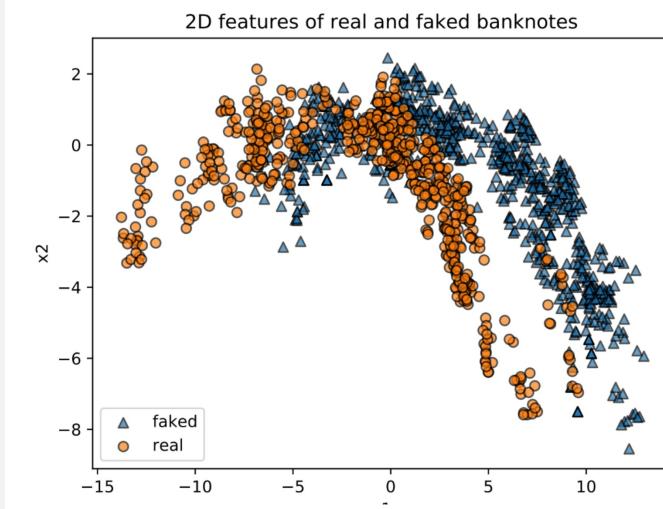


- If training loss does not go down to zero: model is not flexible enough
- In case of overfitting (validation loss >> train loss): regularize model

Discussion of Homework

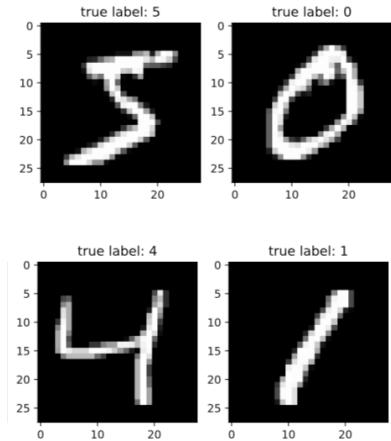
NB 02: Classify banknotes based on 2 features (x_1, x_2)

https://github.com/tensorchiefs/dl_course_2022/blob/master/notebooks/02_fcnn_with_banknote.ipynb



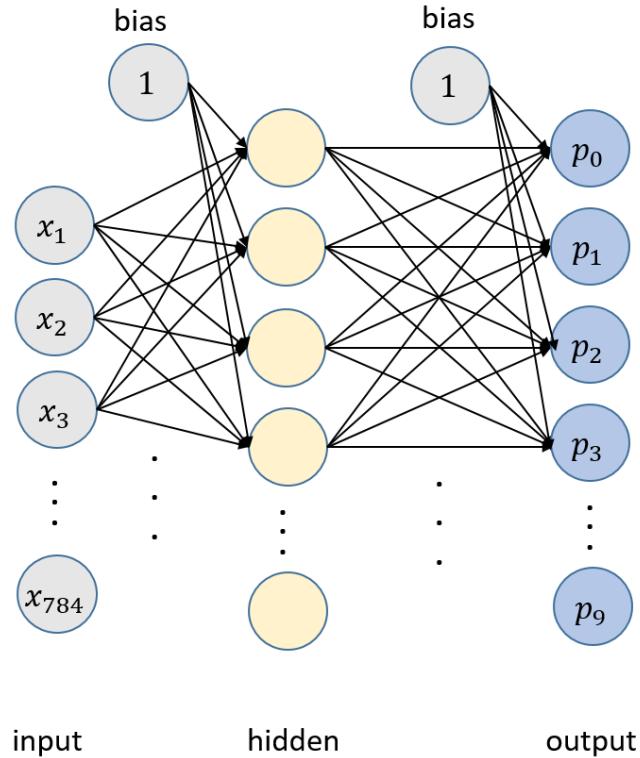
Fully connected NN for image data
Why not?

A fcNN for MNIST data



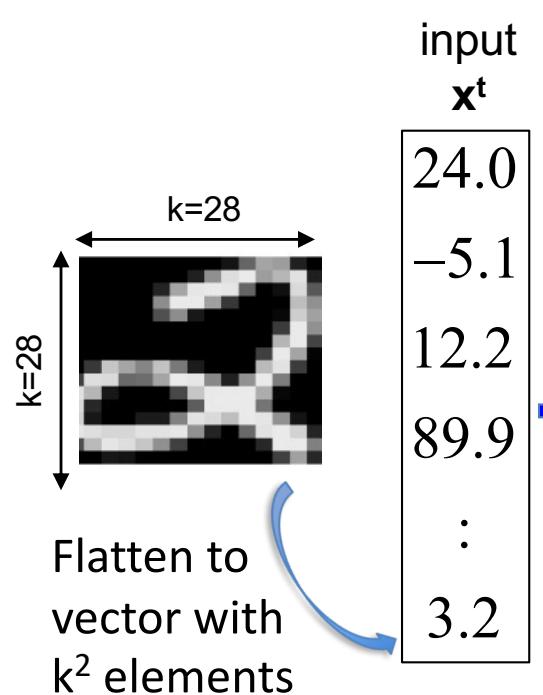
The first four digits of the MNIST data set - each image consisting of $28 \times 28 = 784$ pixels

A fully connected NN with 1 hidden layers.



For the MNIST example, the input layer has 784 values for the 28×28 pixels and the output layer has 10 nodes for the 10 classes (probabilities)

What is going on in a (no hidden layer)



$$\mathbf{x}_{(1,k^2)} \cdot \mathbf{W}_{(k^2,10)} + \mathbf{b}_{(1,10)} = \mathbf{z}_{(1,10)}$$

Score or logit

\mathbf{z}^t

-3.89
-3.18
-0.80
-2.20
-2.44
-1.05
-4.60
-3.48
-2.09
-2.44

$$\hat{p}_k = \frac{e^{z_k}}{\sum_j e^{z_j}}$$

Softmax or probability

$\mathbf{p} = \text{sm}(\mathbf{z})$

0.02
0.04
0.31
0.10
0.08
0.26
0.01
0.03
0.11
0.08

1-hot labels

$obs \mathbf{y}$

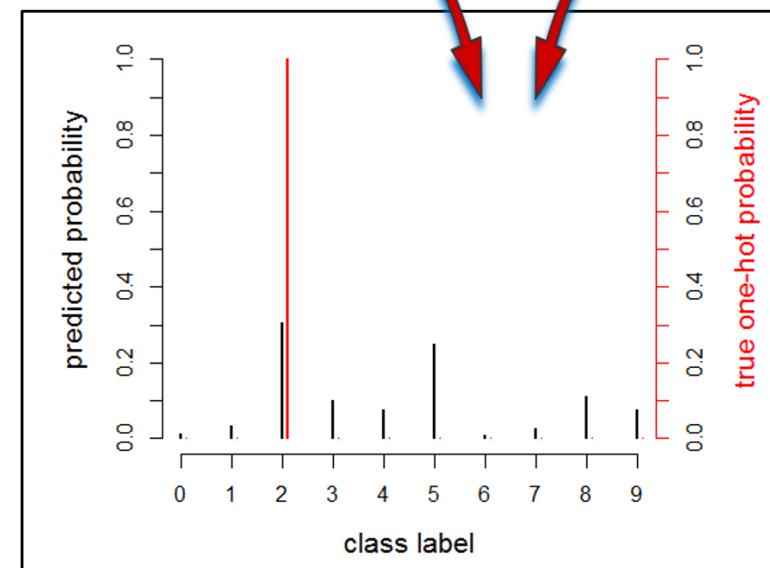
0
0
1
0
0
0
0
0
0
0

Cost C or Loss = cross-entropy averaged over all images in mini-batch

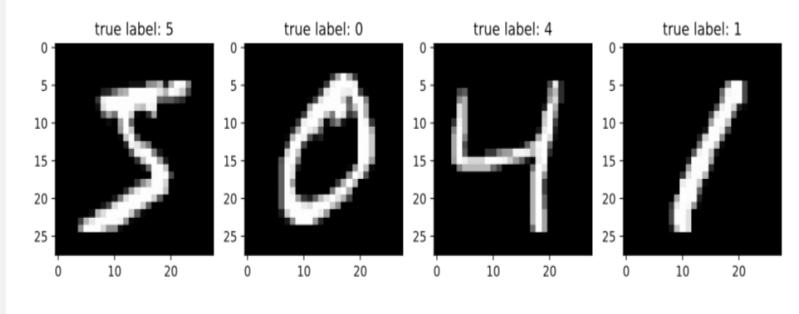
$$C = \frac{1}{N} \sum_i D(\mathbf{p}_i, \mathbf{y}_i)$$

Cross-Entropy

$$D(\mathbf{p}, \mathbf{y}) = -\sum_{k=1}^{10} {}^{obs} y_k \cdot \log(p_k)$$



Exercise: FCNN [1 hour]



	Accuracy on Testset	Loss on Testset
M1 (no hidden)		
M2 (2 hidden)		
M3 (2 hidden ReLU)		

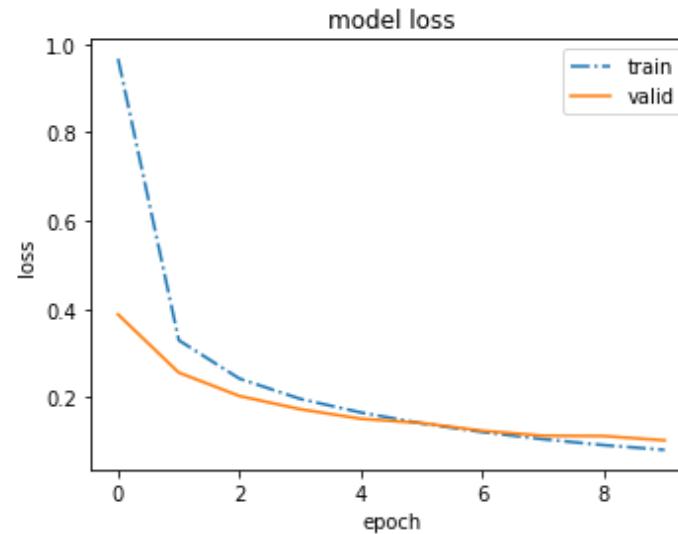
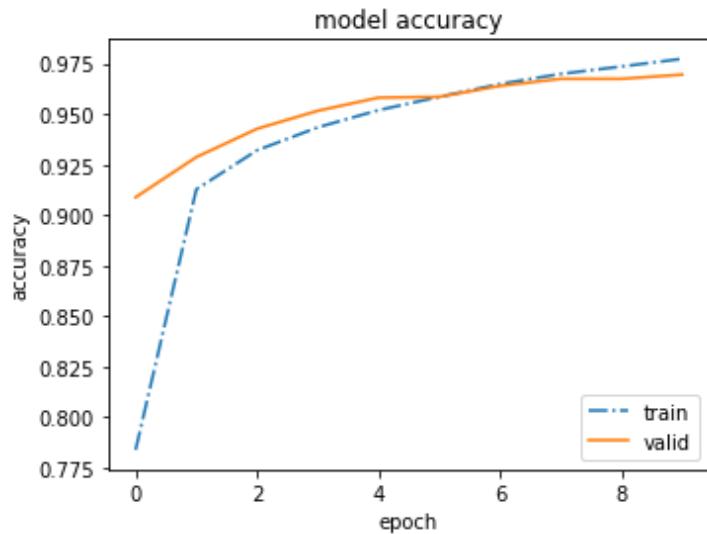
Exercise: fcNN for MNIST 03_fcnn_mnist:

Learning Objective:

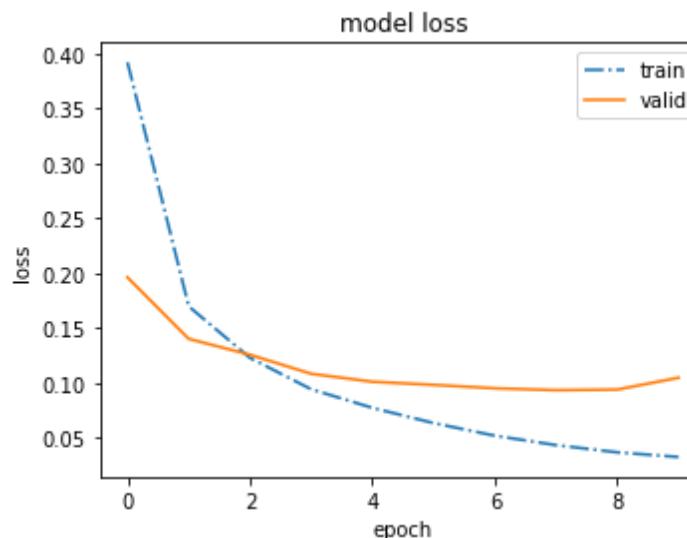
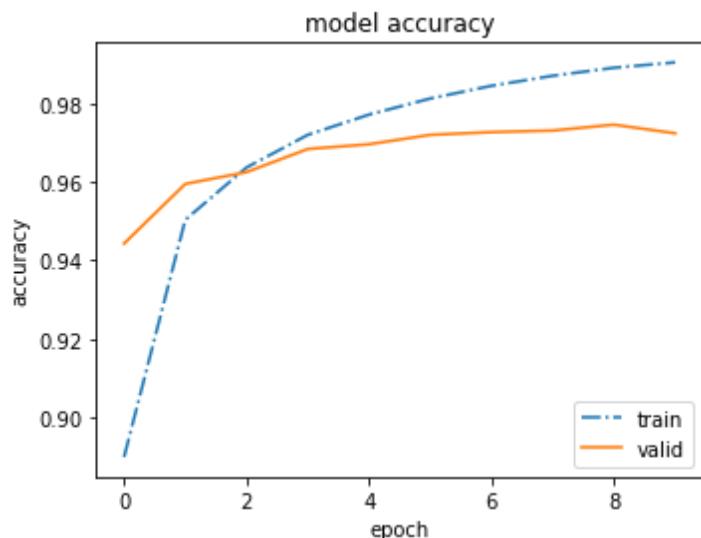
- Effect of ReLU
- Overfitting and Loss-Curves
- Comparing of Models

Result 03_fcNN with ReLU

Sigmoid Acc = 0.9687

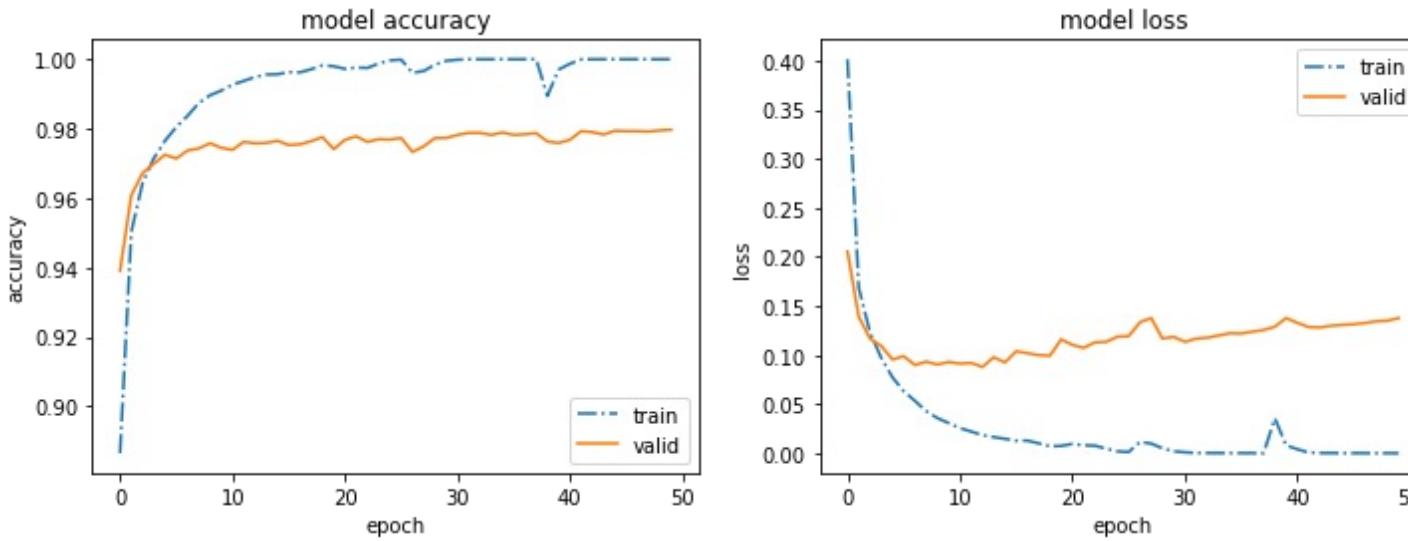


Relu / Acc = 0.9735

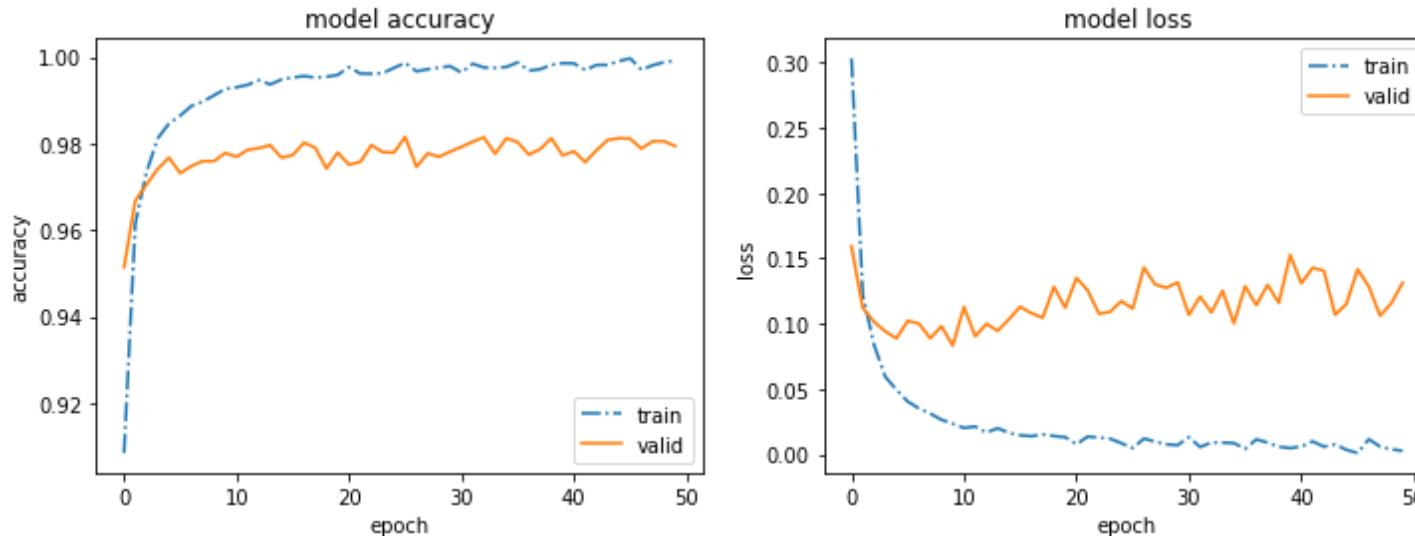


Result 03_fcNN with ReLU longer training

Sigmoid Acc = 0.9789



Relu / Acc = 0.9799



Model comparison

- Model 1 (two hidden Sigmoid)
 - Acc (test-set): 0.9239
- Model 2 (two hidden Sigmoid)
 - Acc (test-set): 0.9687
- Model 3 (two hidden ReLU)
 - Acc (test-set): 0.9735
- Model 4 (two hidden ReLU, long training)
 - Acc (test-set): 0.9799

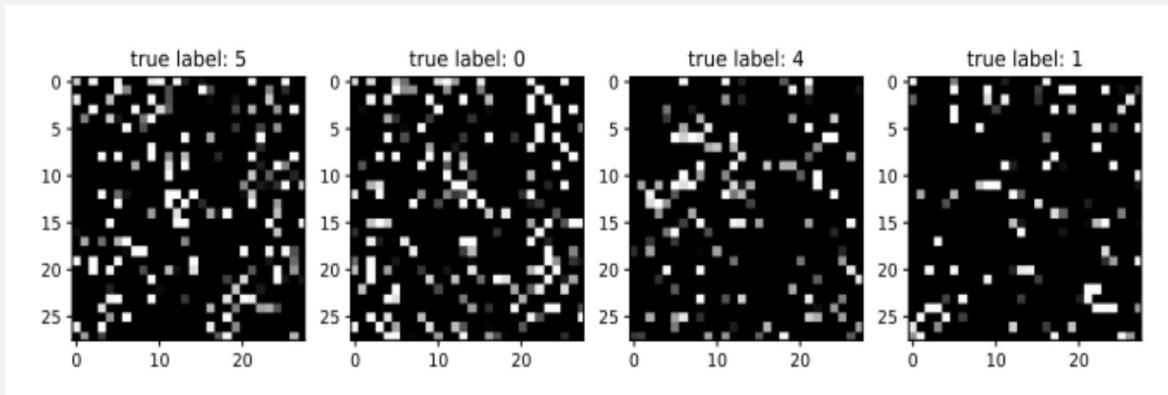
Which model is the best model?

Statistical considerations

It's save to say that model 2 is better than model 1. $M1 < M2$

M3 better than M2, prob. but not 100% sure. Better to compare several runs.

Exercise: FCNN and Shuffling



Exercise : fcNN for MNIST 04_fcnn_mnist

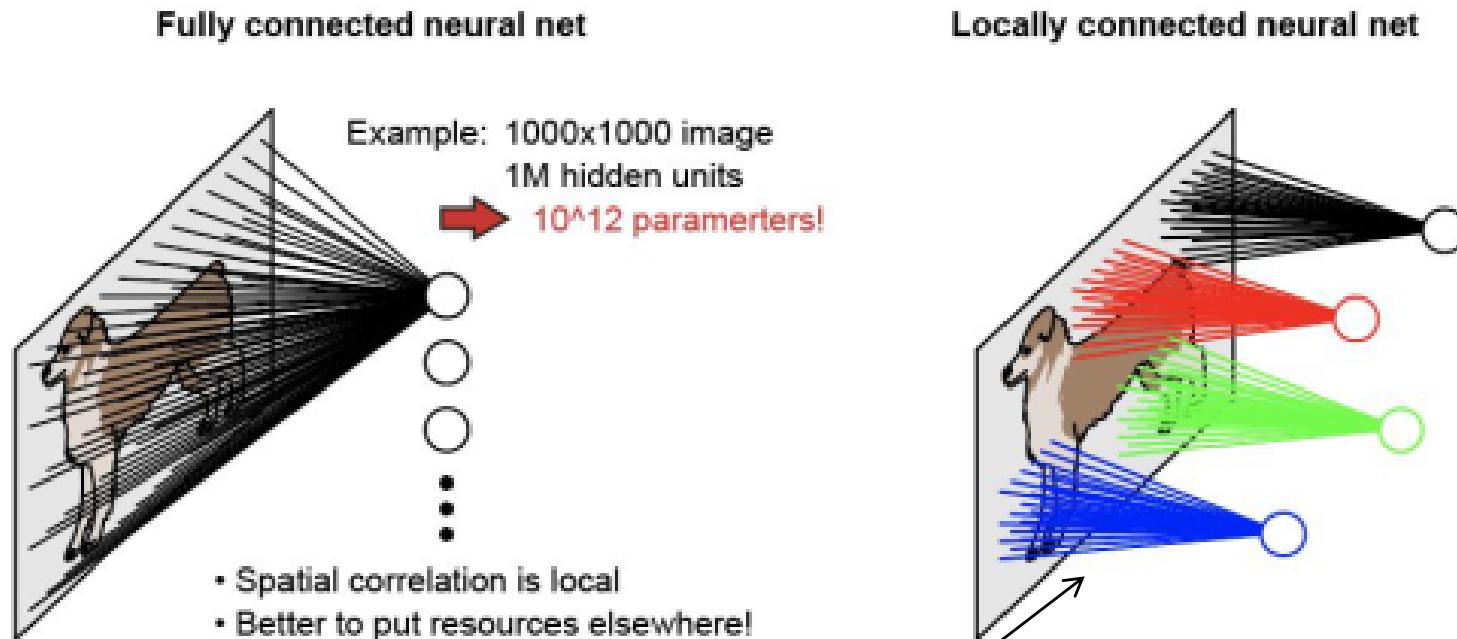
Learning Objective:

Investigate if shuffling disturbs the performance fcNN for MNIST

Convolutional Neural Networks

For image data

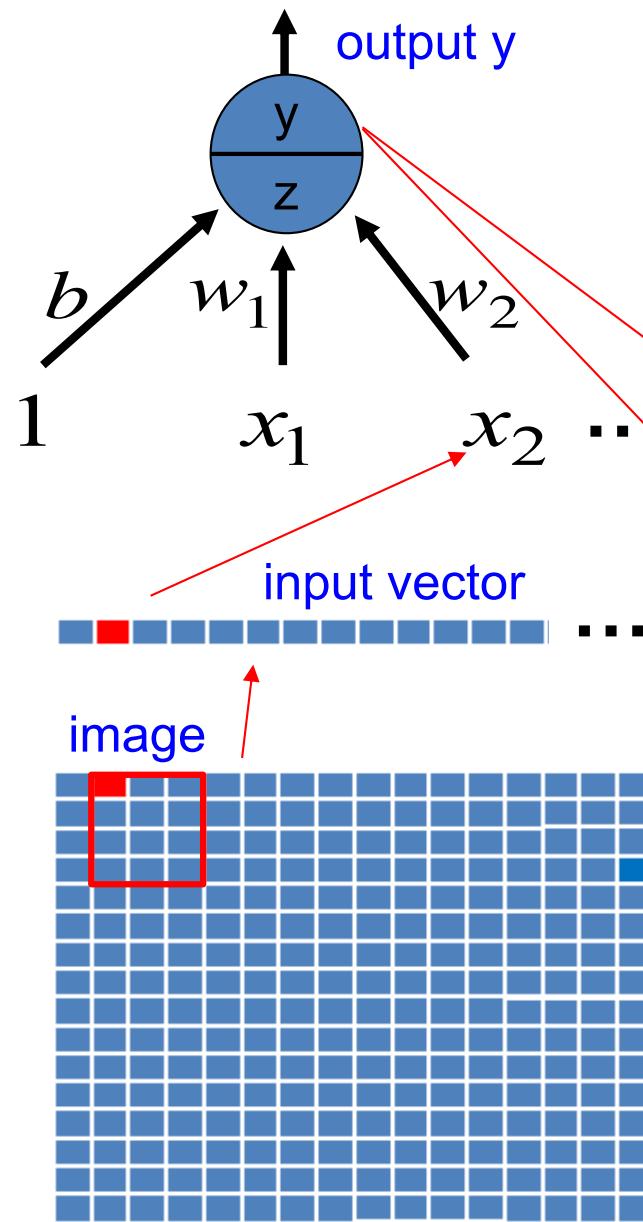
Convolution extracts local information using few weights



Shared weights:

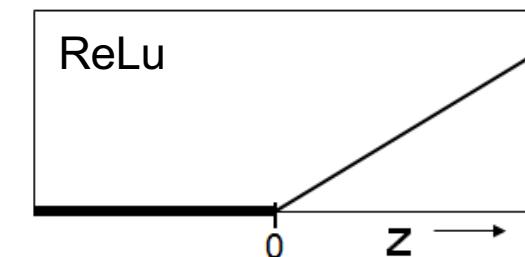
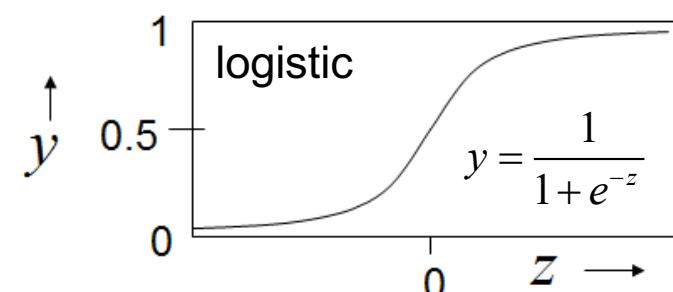
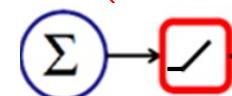
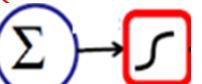
by using the **same weights for each patch** of the image we need much **less parameters** than in the fully connected NN and get from each patch the same kind of **local feature information** such as the presence of an edge.

An artificial neuron



$$z = b + \sum_i x_i w_i$$

Different non-linear transformations are used to get from z to output y

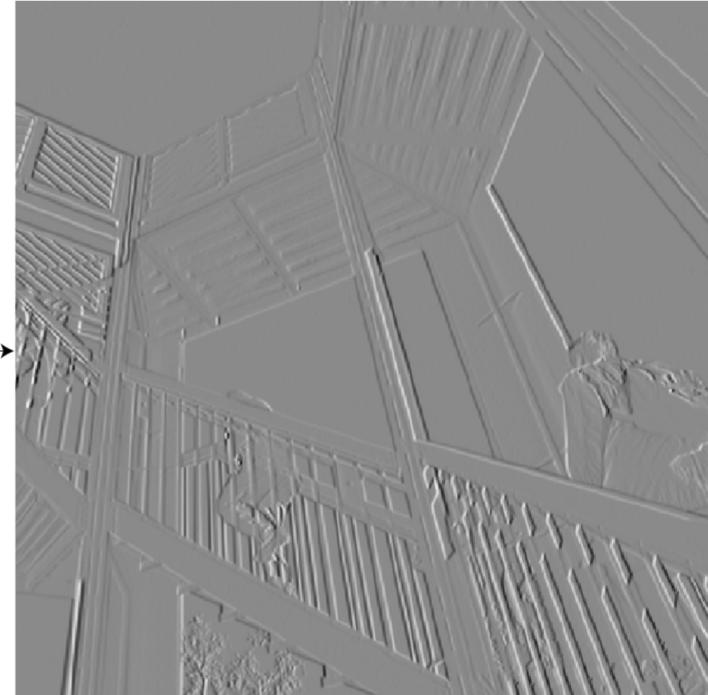


Example of designed Kernel / Filter



$$\begin{bmatrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +1 & 0 & -1 \end{bmatrix}$$

Horizontal Sobel kernel



Applying a vertical edge detector kernel

Taken from the great website on convolution: <https://towardsdatascience.com/intuitively-understanding-convolutions-for-deep-learning-1f6f42faee1>

The basic idea of convolution (blackboard)

2 Dimensional generalization

3 ₀	3 ₁	2 ₂	1	0
0 ₂	0 ₂	1 ₀	3	1
3 ₀	1 ₁	2 ₂	2	3
2	0	0	2	2
2	0	0	0	1

12.0	12.0	17.0
10.0	17.0	19.0
9.0	6.0	14.0

3	3 ₀	2 ₁	1 ₂	0
0	0 ₂	1 ₂	3 ₀	1
3	1 ₀	2 ₁	2 ₂	3
2	0	0	2	2
2	0	0	0	1

12.0	12.0	17.0
10.0	17.0	19.0
9.0	6.0	14.0

3	3	2 ₀	1 ₁	0 ₂
0	0	1 ₂	3 ₂	1 ₀
3	1	2 ₀	2 ₁	3 ₂
2	0	0	2	2
2	0	0	0	1

12.0	12.0	17.0
10.0	17.0	19.0
9.0	6.0	14.0

3	3	2	1	0
0 ₀	0 ₁	1 ₂	3	1
3 ₂	1 ₂	2 ₀	2	3
2 ₀	0 ₁	0 ₂	2	2
2	0	0	0	1

12.0	12.0	17.0
10.0	17.0	19.0
9.0	6.0	14.0

3	3	2	1	0
0	0 ₀	1 ₁	3 ₂	1
3	1 ₂	2 ₂	2 ₀	3
2	0 ₀	0 ₁	2 ₂	2
2	0	0	0	1

12.0	12.0	17.0
10.0	17.0	19.0
9.0	6.0	14.0

3	3	2	1	0
0	0	1 ₀	3 ₁	1 ₂
3	1	2 ₂	2 ₂	3 ₀
2	0	0 ₀	2 ₁	2 ₂
2	0	0	0	1

12.0	12.0	17.0
10.0	17.0	19.0
9.0	6.0	14.0

3	3	2	1	0
0	0	1	3	1
3 ₀	1 ₁	2 ₂	2	3
2 ₂	0 ₂	0 ₀	2	2
2 ₀	0 ₁	0 ₂	0	1

12.0	12.0	17.0
10.0	17.0	19.0
9.0	6.0	14.0

3	3	2	1	0
0	0	1	3	1
3	1 ₀	2 ₁	2 ₂	3
2	0 ₂	0 ₂	2 ₀	2
2	0 ₀	0 ₁	0 ₂	1

12.0	12.0	17.0
10.0	17.0	19.0
9.0	6.0	14.0

3	3	2	1	0
0	0	1	3	1
3	1	2 ₀	2 ₁	3 ₂
2	0	0 ₂	2 ₂	2 ₀
2	0	0 ₀	0 ₁	1 ₂

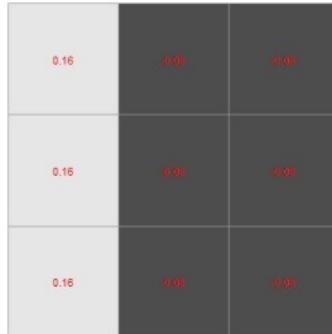
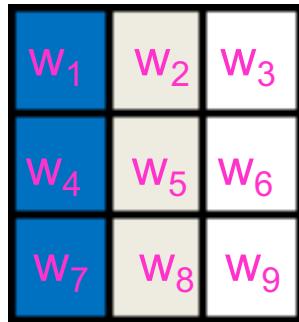
12.0	12.0	17.0
10.0	17.0	19.0
9.0	6.0	14.0

Convolutional networks use neighborhood information and replicated local feature extraction

In a locally connected network the calculation rule

$$z = b + \sum_i x_i w_i$$

Pixel values in a small image patch are element-wise multiplied with weights of a small filter/kernel:

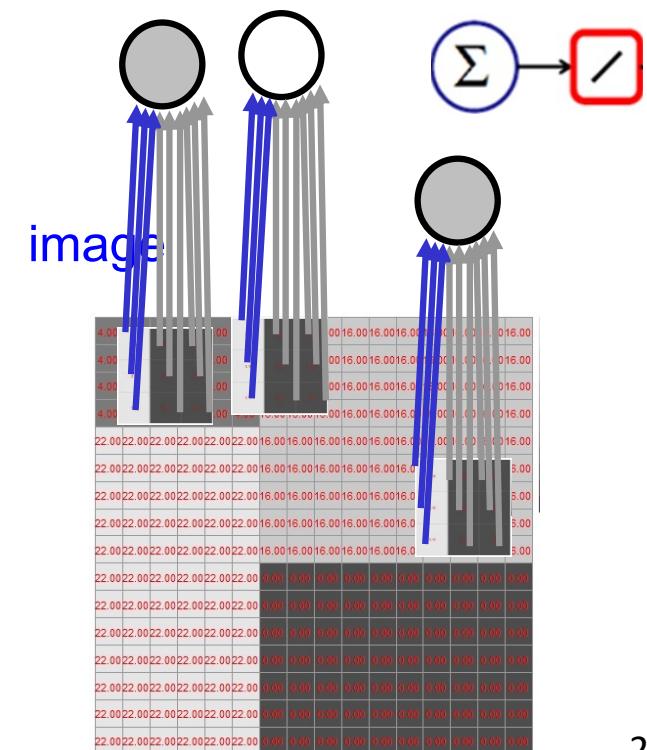


The filter is applied at each position of the image and it can be shown that the result is maximal if the image pattern corresponds to the weight pattern.

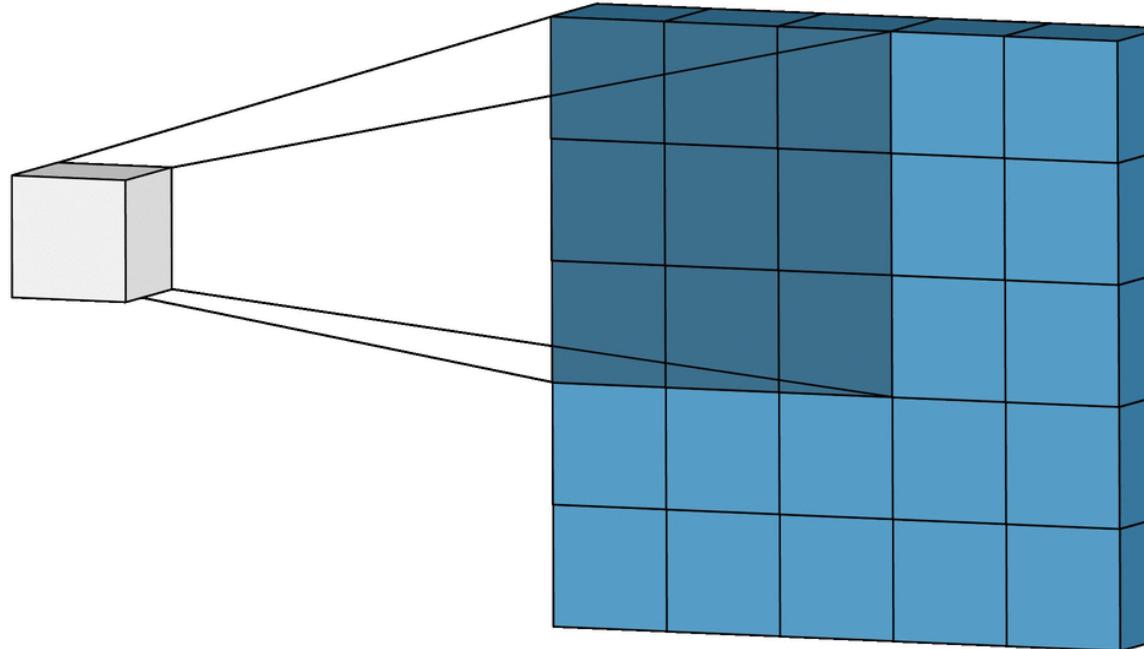
The results form again an image called **feature map** (=activation map) which shows at which position the feature is present.

feature/activation map

-0.00	0.00	0.00	0.00	-0.00	5.82	2.91	0.00	-0.00	-0.00	0.00	0.00	-0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
0.00	0.00	0.00	0.00	-0.00	5.82	2.91	-0.00	0.00	0.00	-0.00	-0.00	0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
0.00	0.00	0.00	0.00	-0.00	5.82	2.91	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
-0.00	0.00	0.00	0.00	-0.00	2.91	1.45	-0.00	0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
0.00	0.00	0.00	0.00	-0.00	0.00	0.00	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
-0.00	0.00	0.00	0.00	-0.00	-2.91	-1.45	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
-0.00	0.00	0.00	0.00	-0.00	-2.91	-1.45	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
0.00	0.00	0.00	0.00	-0.00	-2.91	-1.45	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
0.00	0.00	0.00	0.00	-0.00	-5.82	-2.75	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
0.00	0.00	0.00	0.00	-0.00	8.08	-4.04	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
0.00	0.00	0.00	0.00	-0.00	-10.67	-5.33	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
0.00	0.00	0.00	0.00	-0.00	-10.67	-5.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
0.00	0.00	0.00	0.00	-0.00	-10.67	-5.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
0.00	0.00	0.00	0.00	-0.00	-10.67	-5.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
0.00	0.00	0.00	0.00	-0.00	-10.67	-5.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
0.00	0.00	0.00	0.00	-0.00	-10.67	-5.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
0.00	0.00	0.00	0.00	-0.00	-10.67	-5.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00



Applying the same 3×3 kernel at each image position



Applying the 3×3 kernel on a certain position of the image yields one pixel within the activation map where the position corresponds to the center of the image patch on which the kernel is applied.

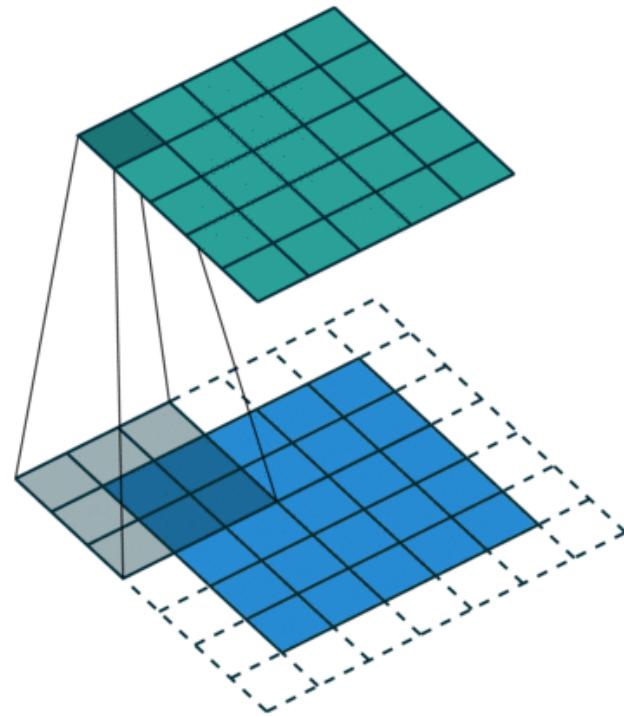
Exercise: Do one convolution step by hand



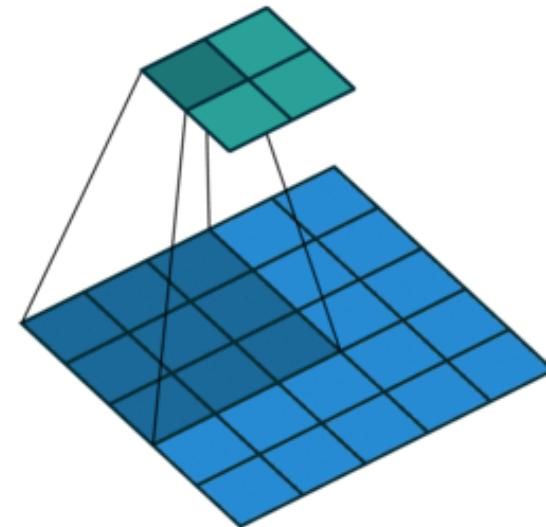
1. The kernel is 3x3 and is applied at each valid position
– how large is the resulting activation map?
2. The small numbers in the shaded region are the kernel weights.
Determine the position and the value within the resulting activation map.

3	3	2	1	0
0 ₀	0 ₁	1 ₂	3	1
3 ₂	1 ₂	2 ₀	2	3
2 ₀	0 ₁	0 ₂	2	2
2	0	0	0	1

Padding



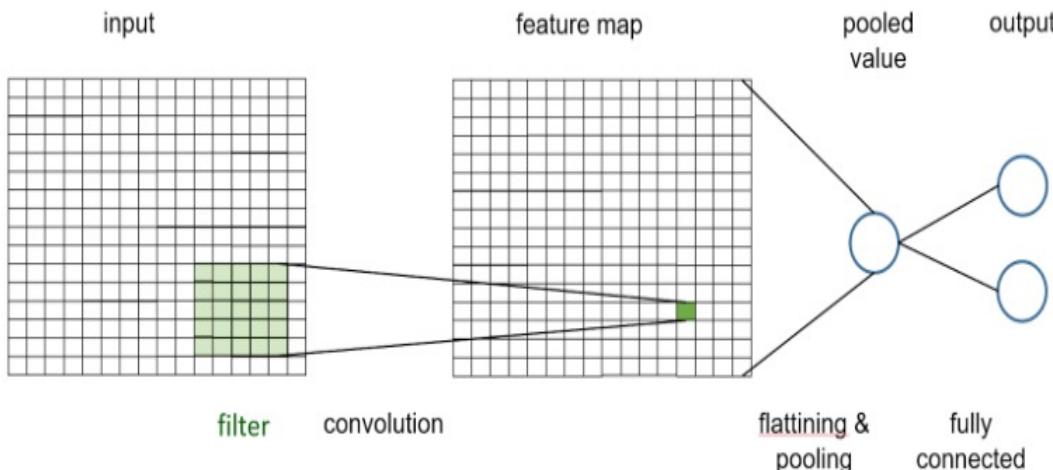
Zero-padding to achieve
same size of feature and input



no padding to only use
valid input information

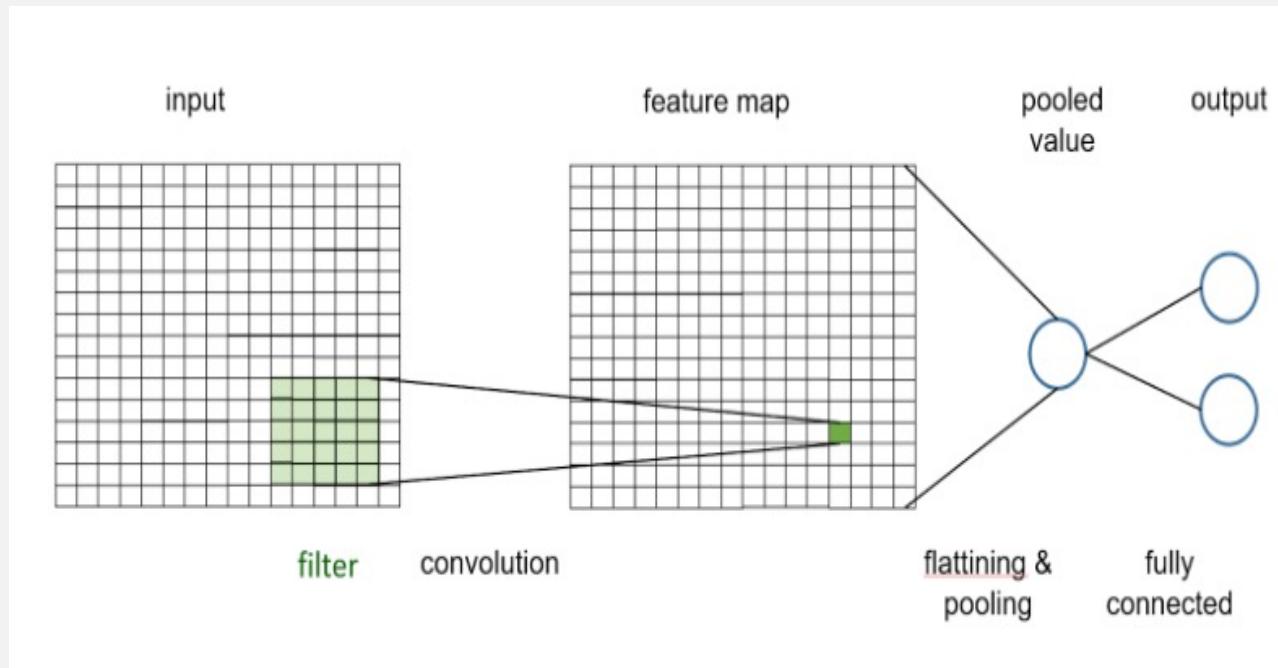
The *same* weights are used at each position of the input image.

Building a very simple CNN with keras



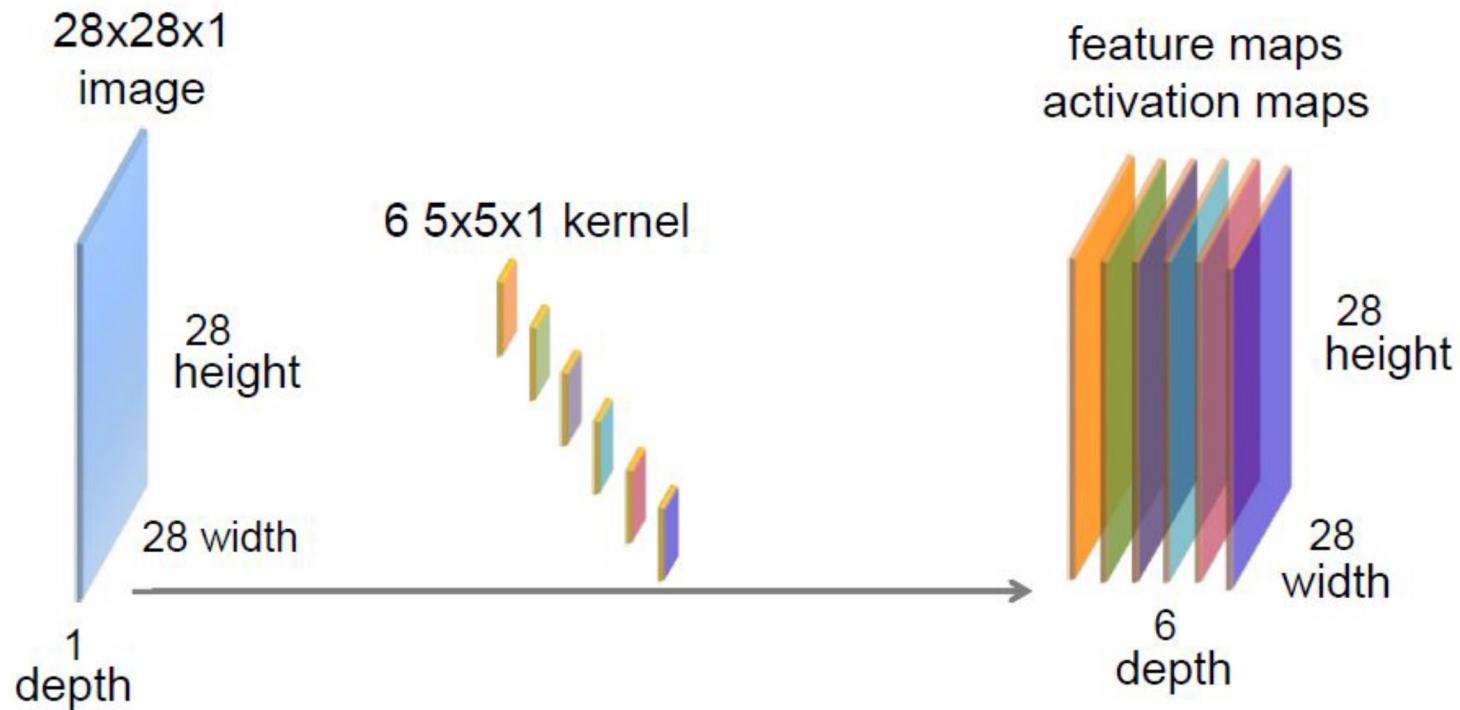
```
model = Sequential()
model.add(Convolution2D(1, (5,5), # one 5x5 kernel
                      padding='same',           # zero-padding to preserve size
                      input_shape=(pixel,pixel,1)))
model.add(Activation('linear'))
# take the max over all values in the activation map
model.add(MaxPooling2D(pool_size=(pixel,pixel)))
model.add(Flatten())
model.add(Dense(2))
model.add(Activation ('softmax'))
```

Exercise: Artstyle Lover



Open NB 05_cnn_edge_lover.ipynb

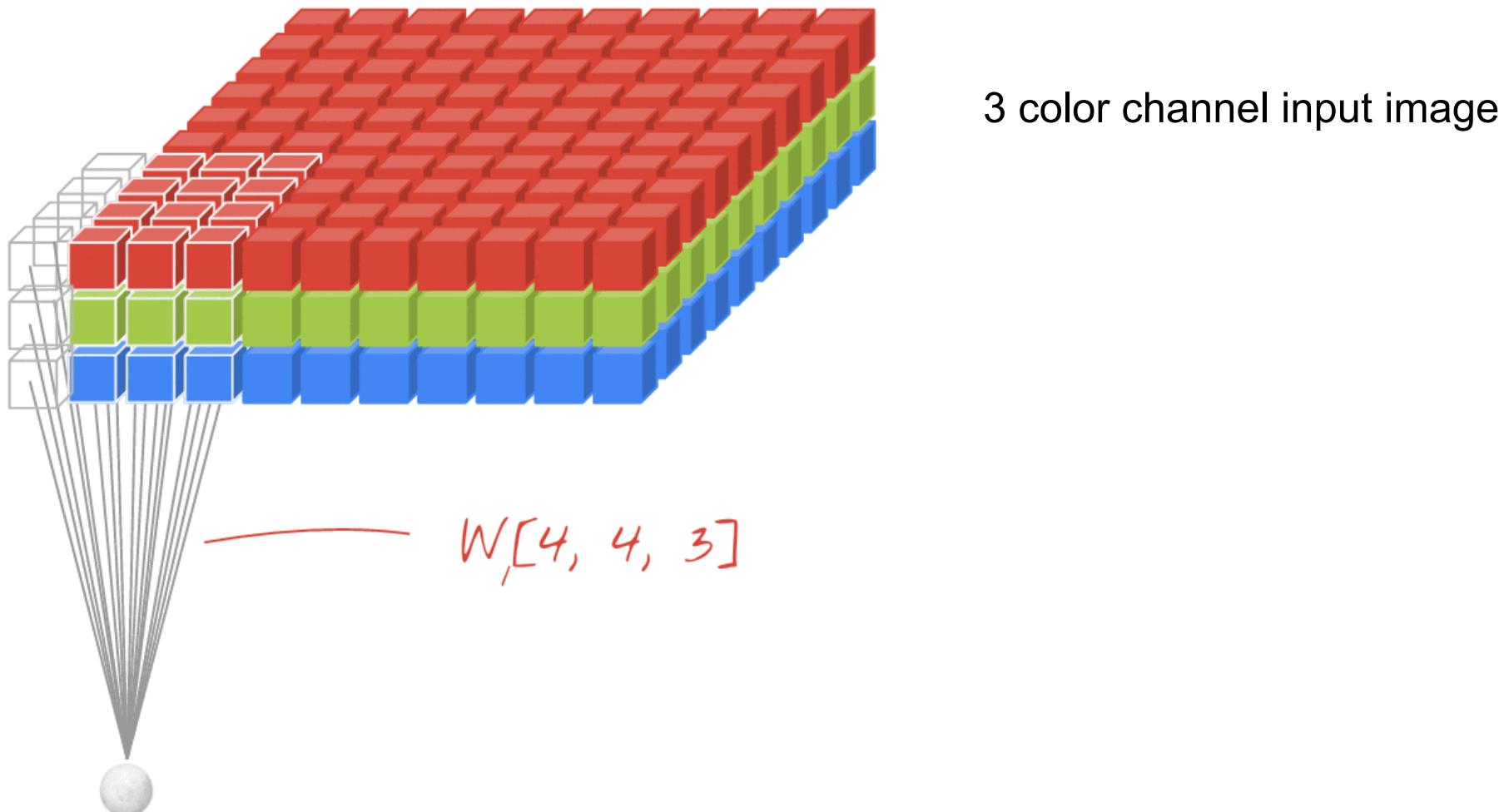
Convolution layer with a 1-channel input and 6 kernels



Convolution of the input image with 6 different kernels results in 6 activation maps.

If the input image has only one channel, then each kernel has also only one channel.

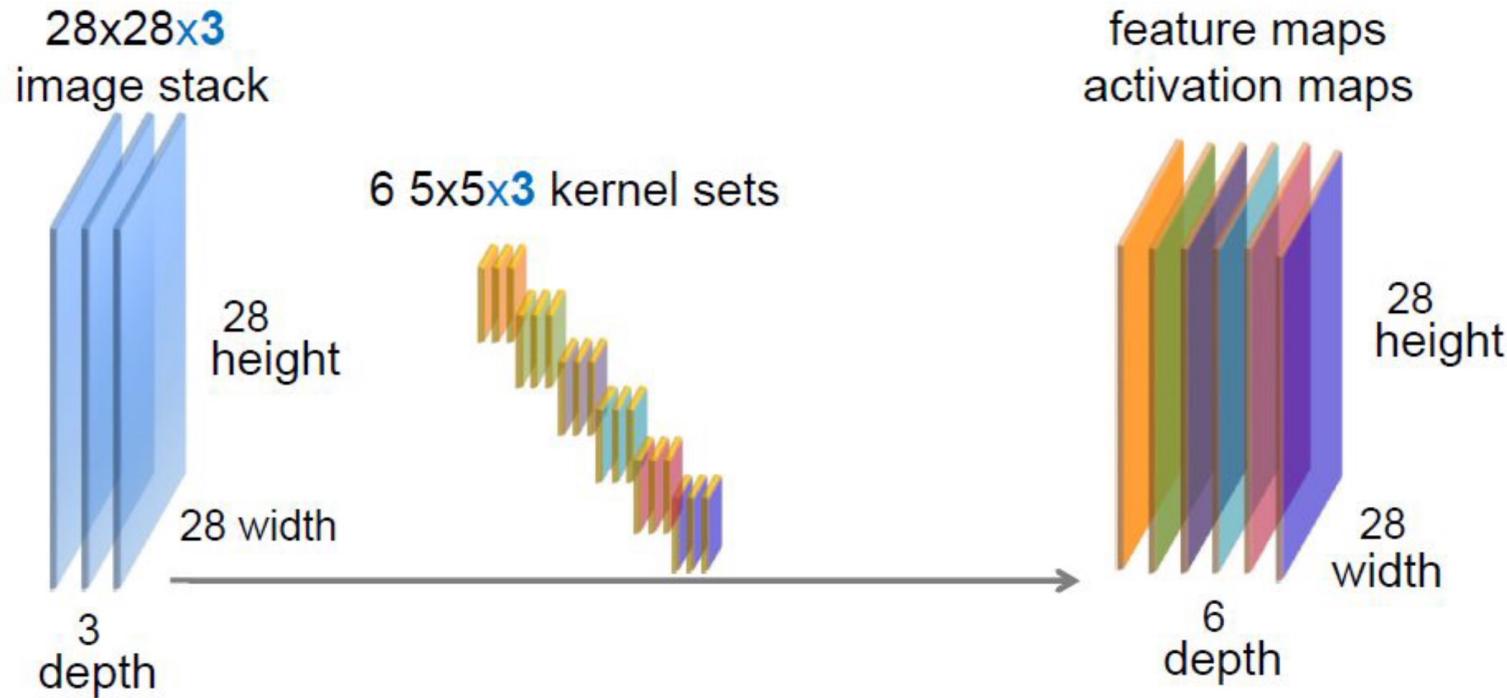
Animated convolution with 3 input channels



The value of neuron j in the k -th featuremap are computed from the weights in the k -th filter w_{ki} and the input values x_{ji} at the position j :

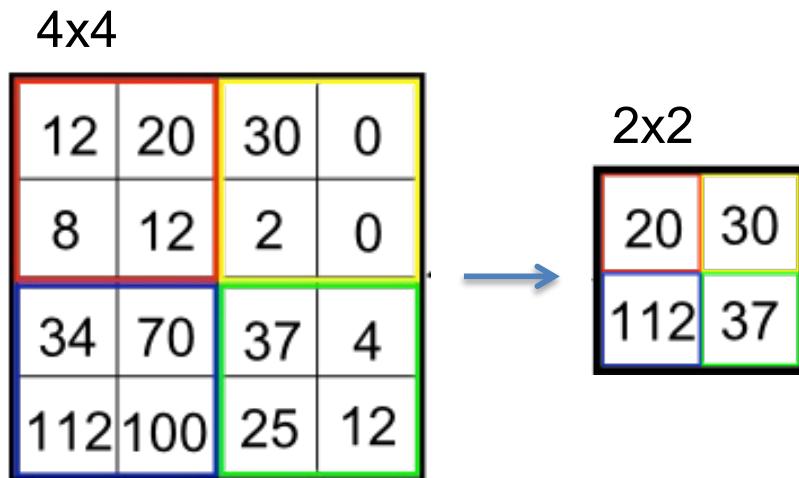
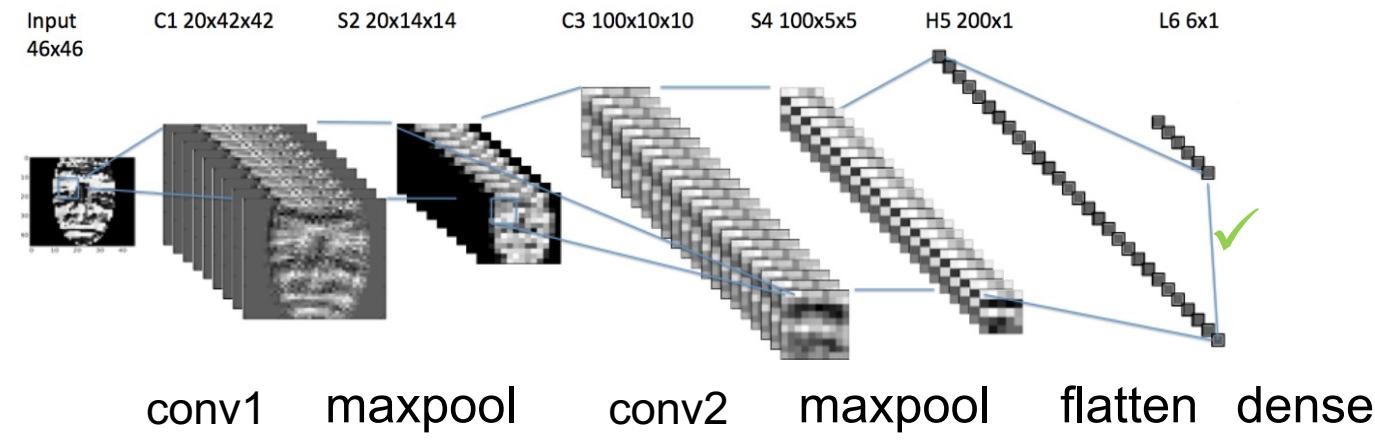
$$y_{jF_k} = f(z_{jF_k}) = f(b_k + \sum x_{ji} \cdot w_{ki})$$

Convolution layer with a 3-chanel input and 6 kernels



Convolution of the input image with 6 different kernels results in 6 activation maps.
If the input image has 3 channels, then each filter has also 3 channels.

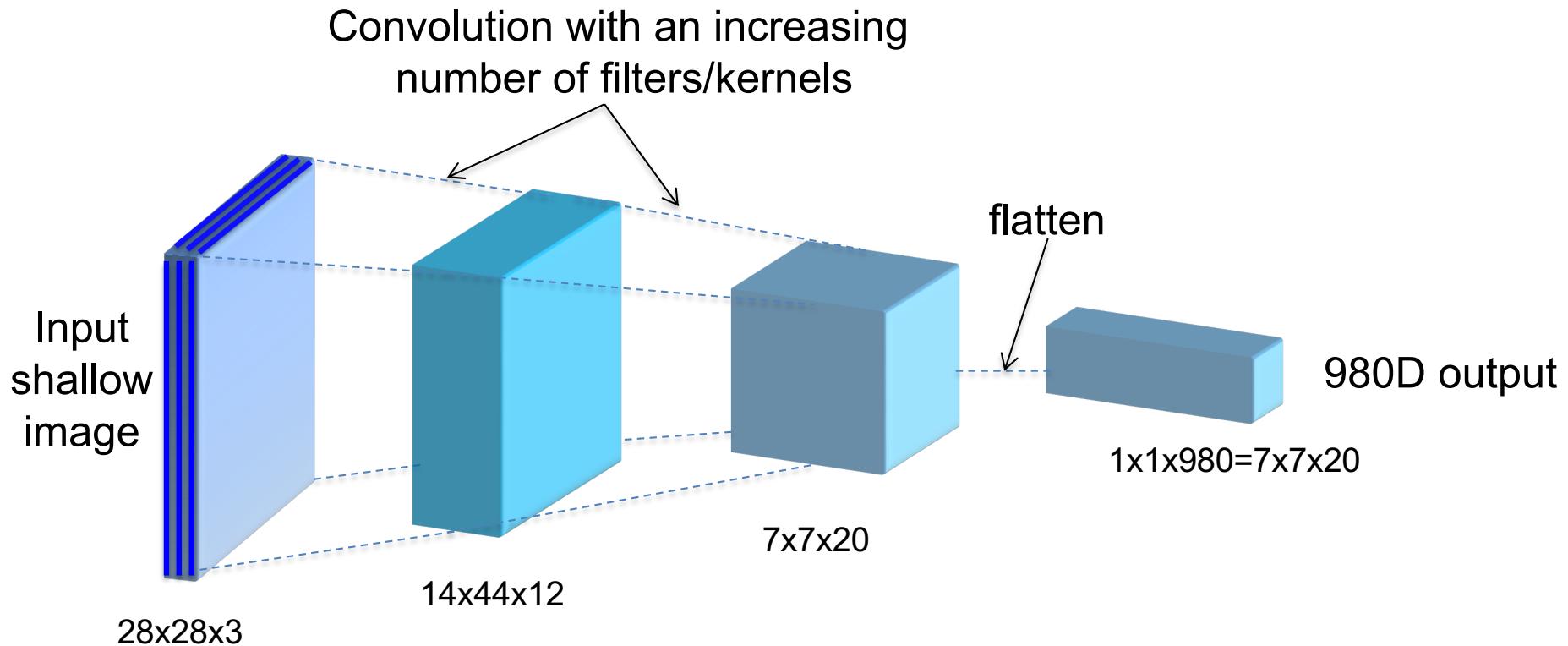
CNN ingredient II: Maxpooling Building Blocks reduce size



Simply join e.g. 2x2 adjacent pixels in one by taking the max.
→ less weights in model
→ Less train data needed
→ increased performance

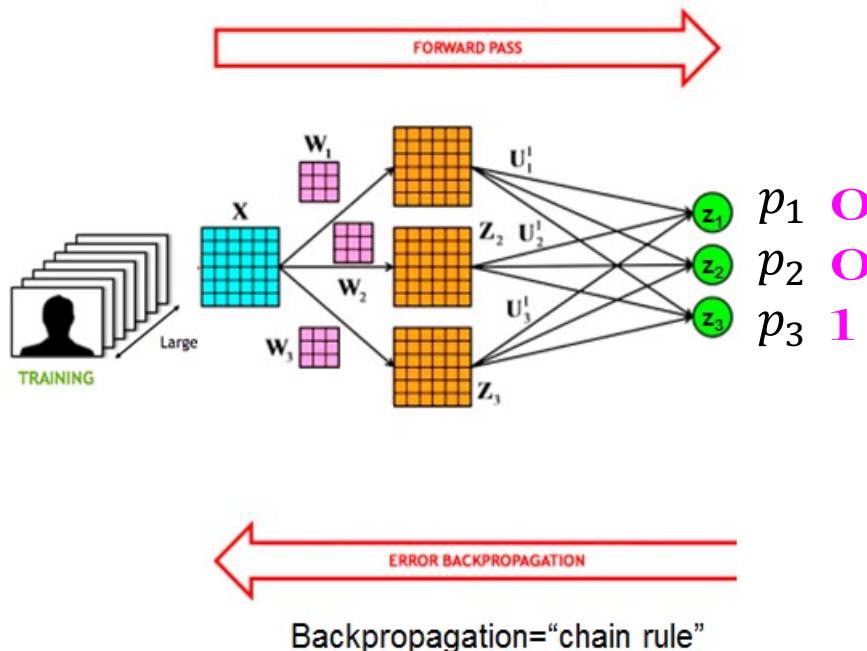
Hinton: „The pooling operation used in convolutional neural networks is a big mistake and the fact that it works so well is a disaster“

Typical shape of a classical CNN



Spatial resolution is decreased e.g. via max-pooling while more abstract image features are detected in deeper layers.

Training of a CNN is based on gradient backpropagation



Learning is done by weight updating:

For the training we need the **observed label** for each image which we then compare with the **output** of the CNN.

We want to **adjust the weights** in a way so that difference between true label and output is minimal.

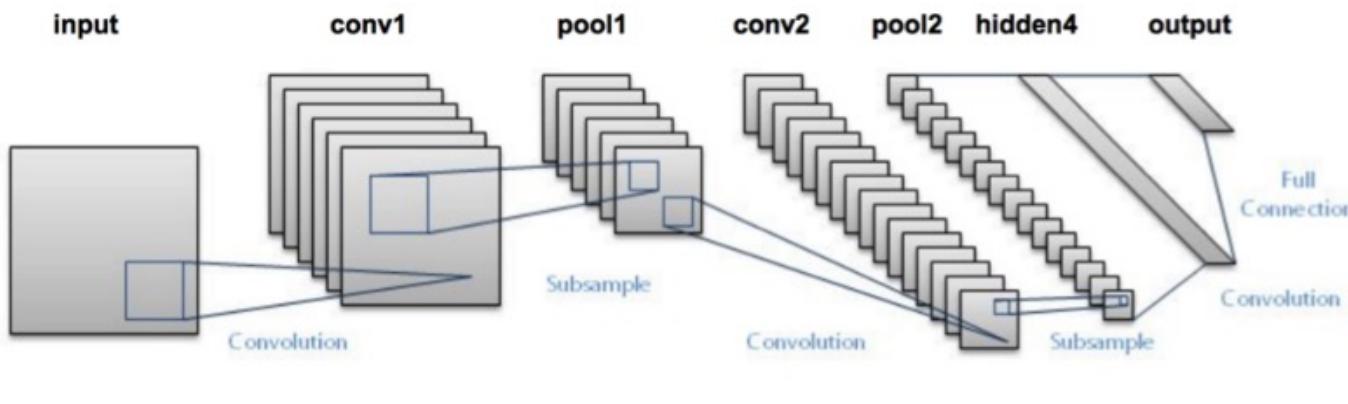
Minimize Loss-function:

$$L = \text{distance}(\text{observed}, \text{output}(w))$$

$$w_i^{(t)} = w_i^{(t-1)} - l^{(t)} \left. \frac{\partial L(w)}{\partial w_i} \right|_{w_i=w_i^{(t-1)}}$$

learning rate

CNN for MNIST



```
num_classes = 10                      # 10 classes: 0, 1,...,9
input_shape = (28, 28, 1)    # 28x28 pixels, 1 channel (grey value)

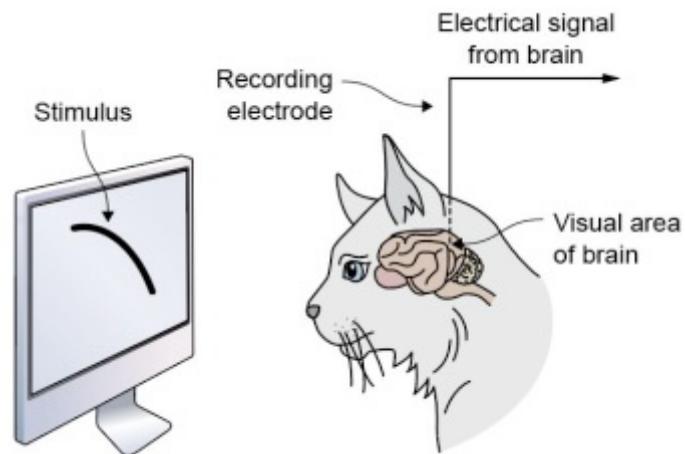
model = Sequential()
model.add(Convolution2D(32, (3, 3), #32 filters of size 3x3
                      activation='relu',
                      input_shape=input_shape))
model.add(MaxPooling2D(pool_size=(2, 2)))
model.add(Convolution2D(64, (3, 3), activation='relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))
model.add(Flatten())
model.add(Dense(40, activation='relu'))
model.add(Dense(num_classes, activation='softmax'))
```

fcNN versus CNNs – some aspects

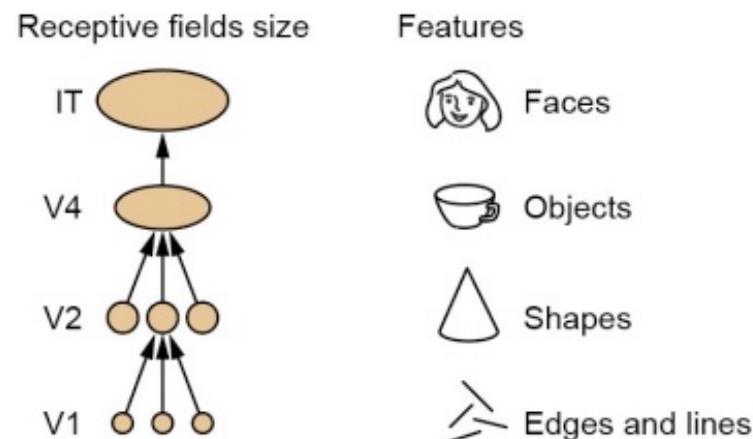
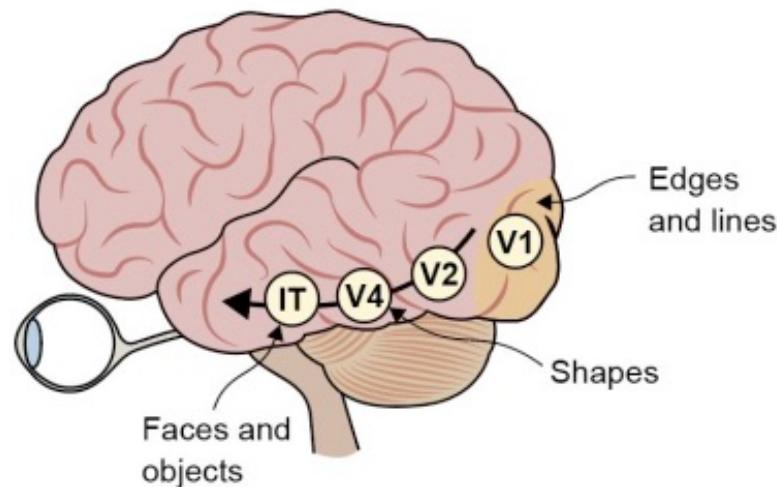
- A fcNN is good for tabular data, CNNs are good for ordered data (eg images).
- In a fcNN the order of the input does not matter, in CNN shuffling matters.
- A fcNN has no model bias, a CNN has the model bias that neighborhood matters.
- A node in one layer of a fcNN corresponds to one feature map in a convolution layer:
- In each layer of a fcNN connecting p to q nodes, we learn q linear combinations of the incoming p signals, in each layer of a CNN connecting p channels with q channels we learn q filters (each having p channels) yielding q feature maps

Biological Inspiration of CNNs

How does the brain respond to visually received stimuli?



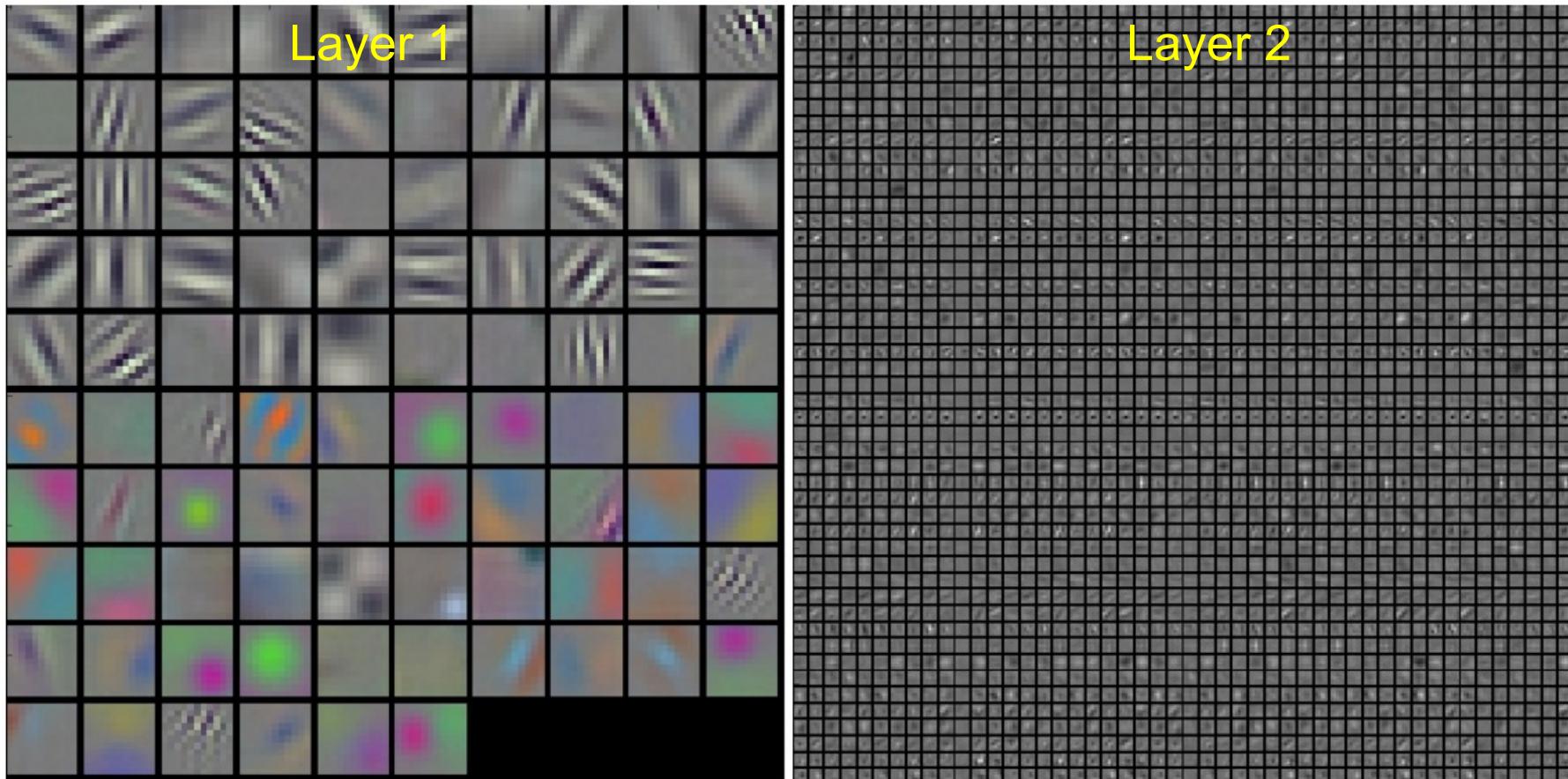
Setup of the experiment of Hubel and Wiesel in late 1950s in which they discovered **neurons** in the visual cortex that **responded** when moving **edges** were shown to the cat.



Organization of the visual cortex in a brain, where neurons in different regions respond to more and more complex stimuli

Visualize the weights used in filters

Filter weights from a trained Alex Net

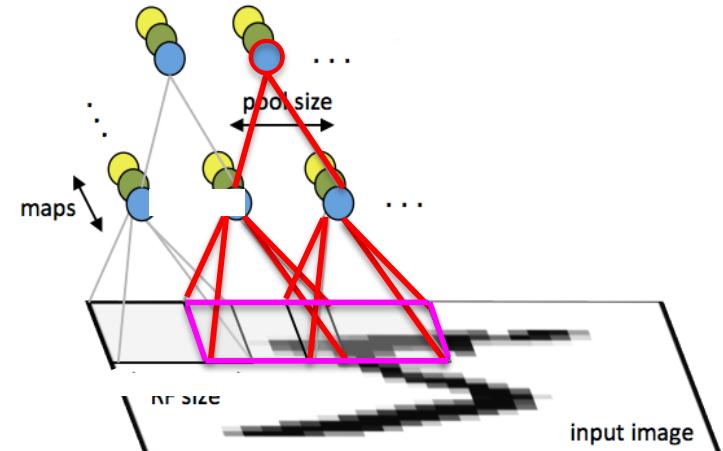
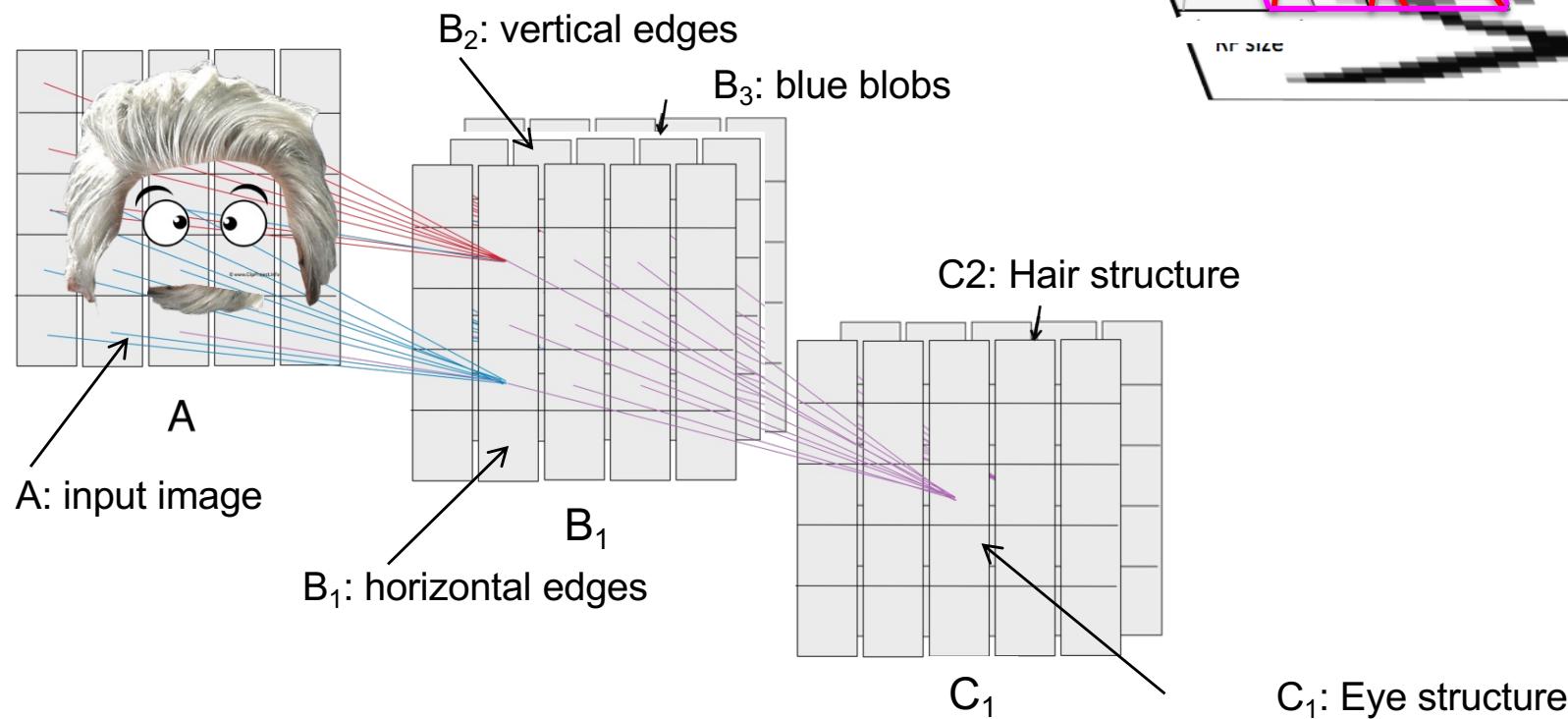


Only in layer 1 the filter pattern correspond to extracted patterns in the image.

In higher layers we can only check if patterns look noisy, which would indicate that the network that hasn't been trained for long enough, or possibly with a too low regularization strength that may have led to overfitting.

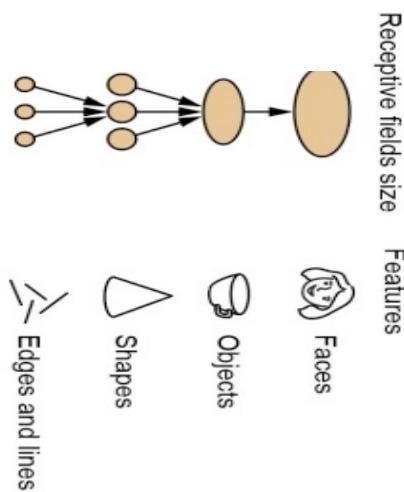
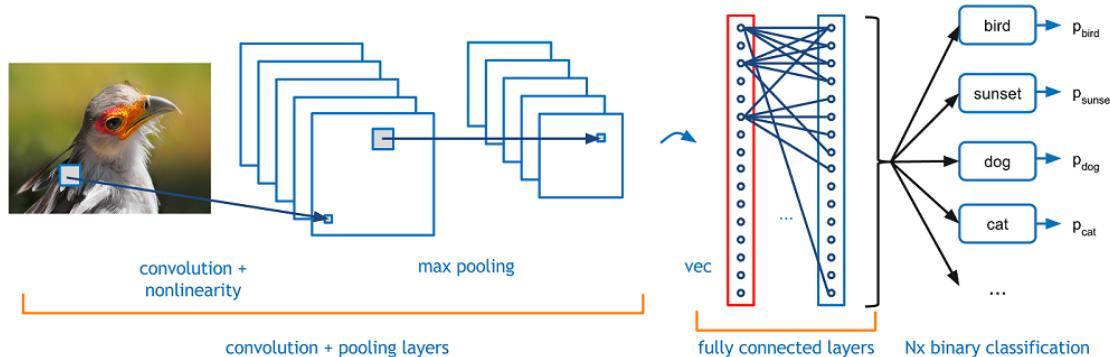
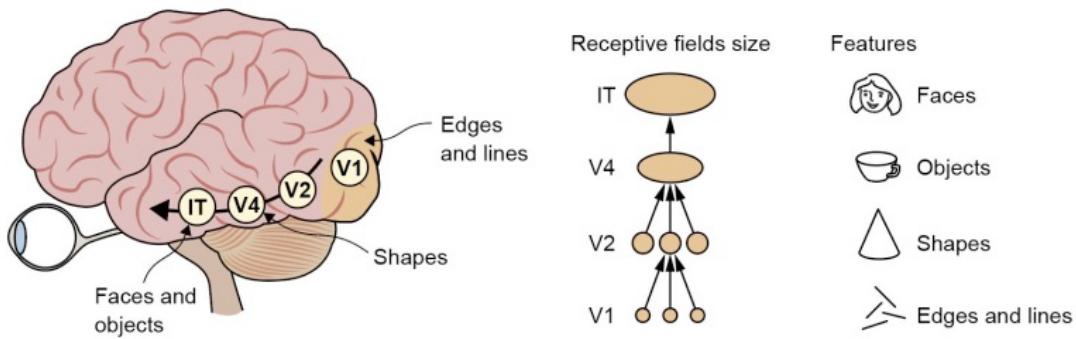
The receptive field

For each pixel of a feature map we can determine the connected area in the input image – this area in the input image is called receptive field.

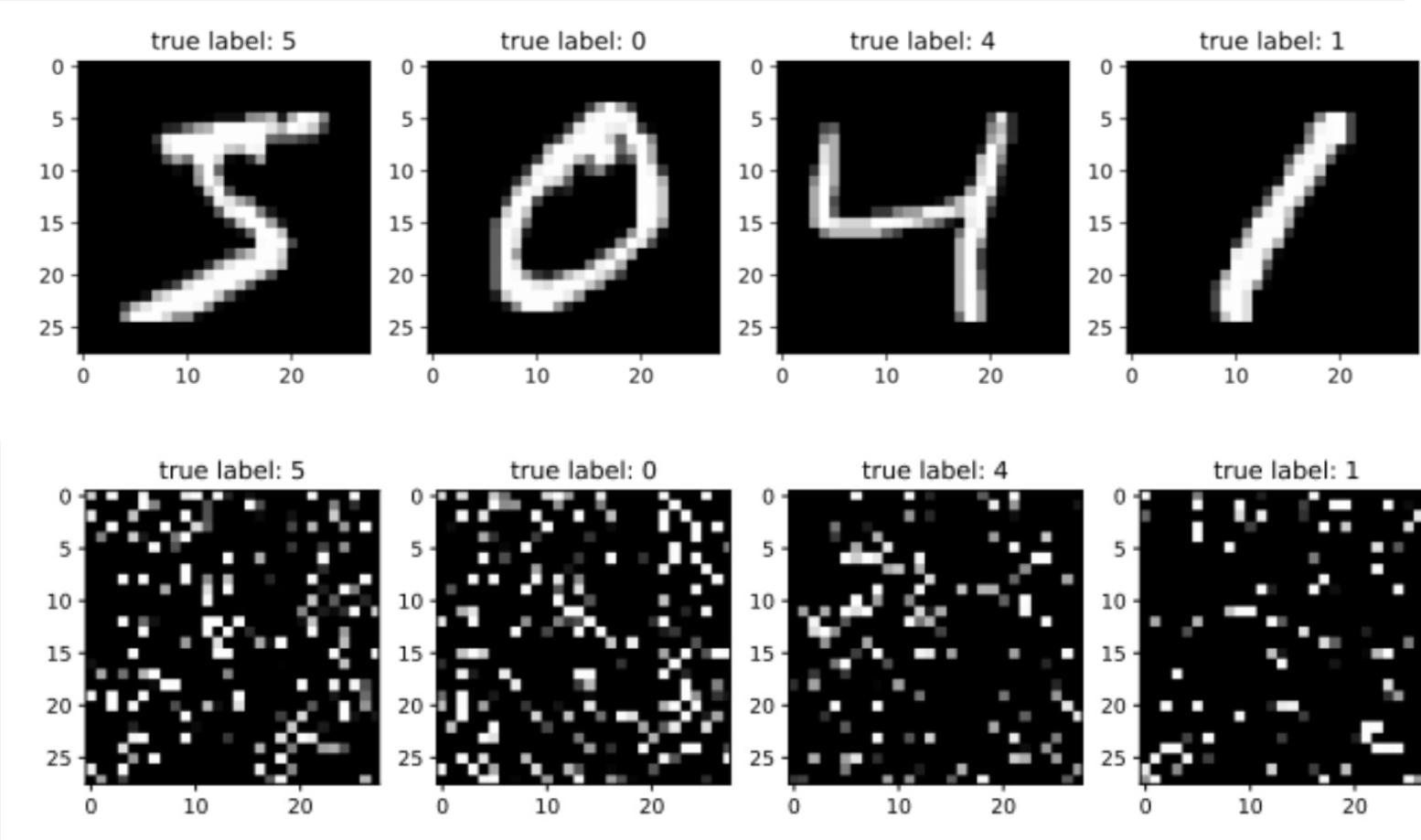


A feature map gets activated by a certain structure of the feature maps one layer below, which by itself depends on the input of a preceding layer etc and finally on the input image. Activation maps close to the input image are activated by simple structures in the image, higher maps by more complex image structures. 41

Weak analogies between brain and CNNs architecture



Does shuffling disturb a CNN?



Inverstigate if shuffling disturbs the CNN for MNIST **06_cnn_mnist_shuffled**

Summary

- Use loss curves to detect overfitting or underfitting problems
- NNs work best when respecting the underlying structure of the data.
 - Use fully connected NN for tabular data
 - Use convolutional NN for data with local order such as images
- CNNs exploit the local structure of images by local connections and shared weight (same kernel is applied at each position of the image).
- Use the relu activation function for hidden layers in CNNs.
- NNs are loosely inspired by the structure of the brain.
 - When going deep the receptive field increases (~layer 5 sees whole input)
 - Deeper layer respond to more complex feature in the input