AARHUS
UNIVERSITET

Convolutional Neural Networks

Henrik Karstoft
@
Aarhus University, Department of Engineering
hka@eng.au.dk

Convolutional Neural Networks uses

Image classification
- Coarse (high-level objects)
- Fine grained (dog, bird weed species) harder problem

- Object detection
  - Bounding box regression, YOLO

- Image segmentation
  - Fully-connected networks
  - U architectures

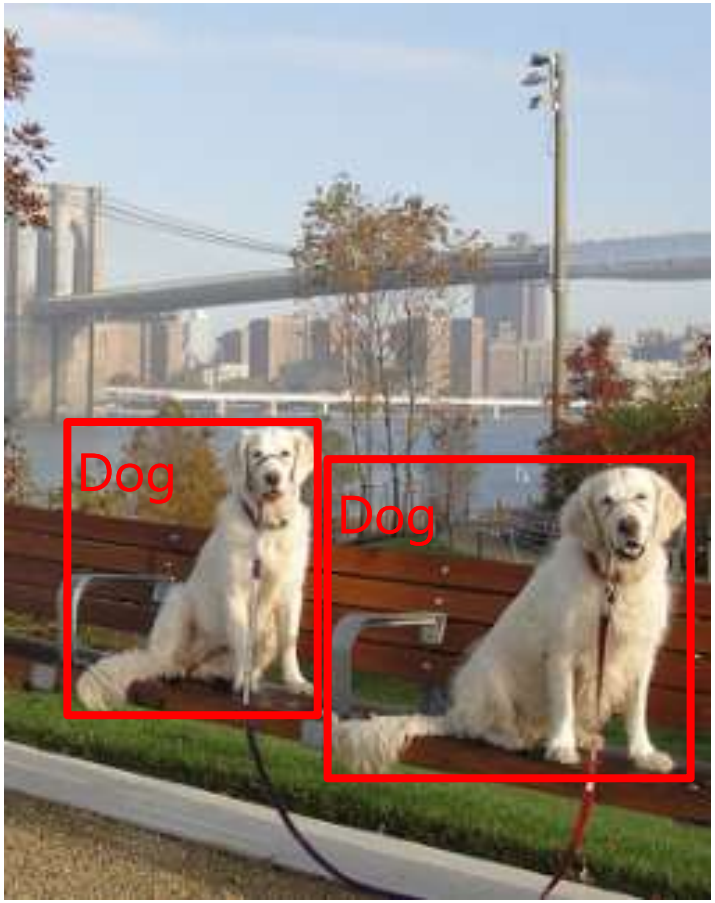
- Synthesis and visualization
  - Adversarial networks Simon

- Sentence generation
  - Recurrent CNNs
  - LSTMs

- Depth-map estimation

- ...

Classification and Detection

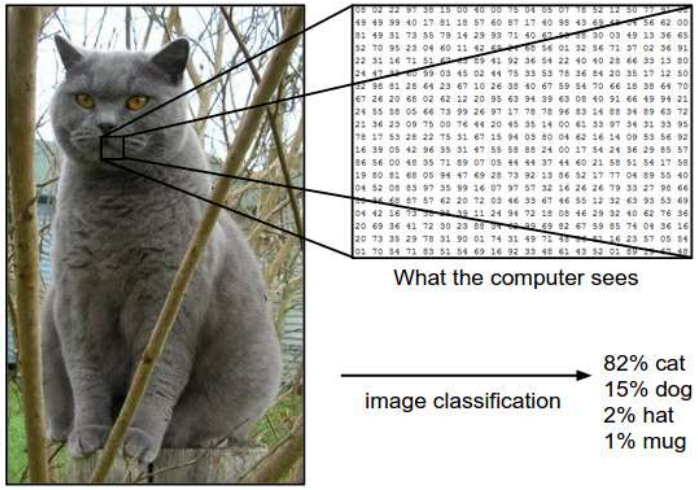


Classification

Visual Object classification

Challenges

- **Viewpoint**
- **Scale**
- **Deformation**
- **Occlusion**
- **Illumination conditions**
- **Intra-class variation**



Classification

The classical pattern classification mod… 50's)

Hand-crafted
Feature extractor

Simple Trainable
Classifier

…lassic car"

Hard and art and time consuming

Less hard

Builds Data Representations
on increasing abstraction levels

Modern approach taking off in 2006 (Hinton et al.) an… 12 (Alexnet)
(feature learning and classifier learning)

Trainable
Feature extractor

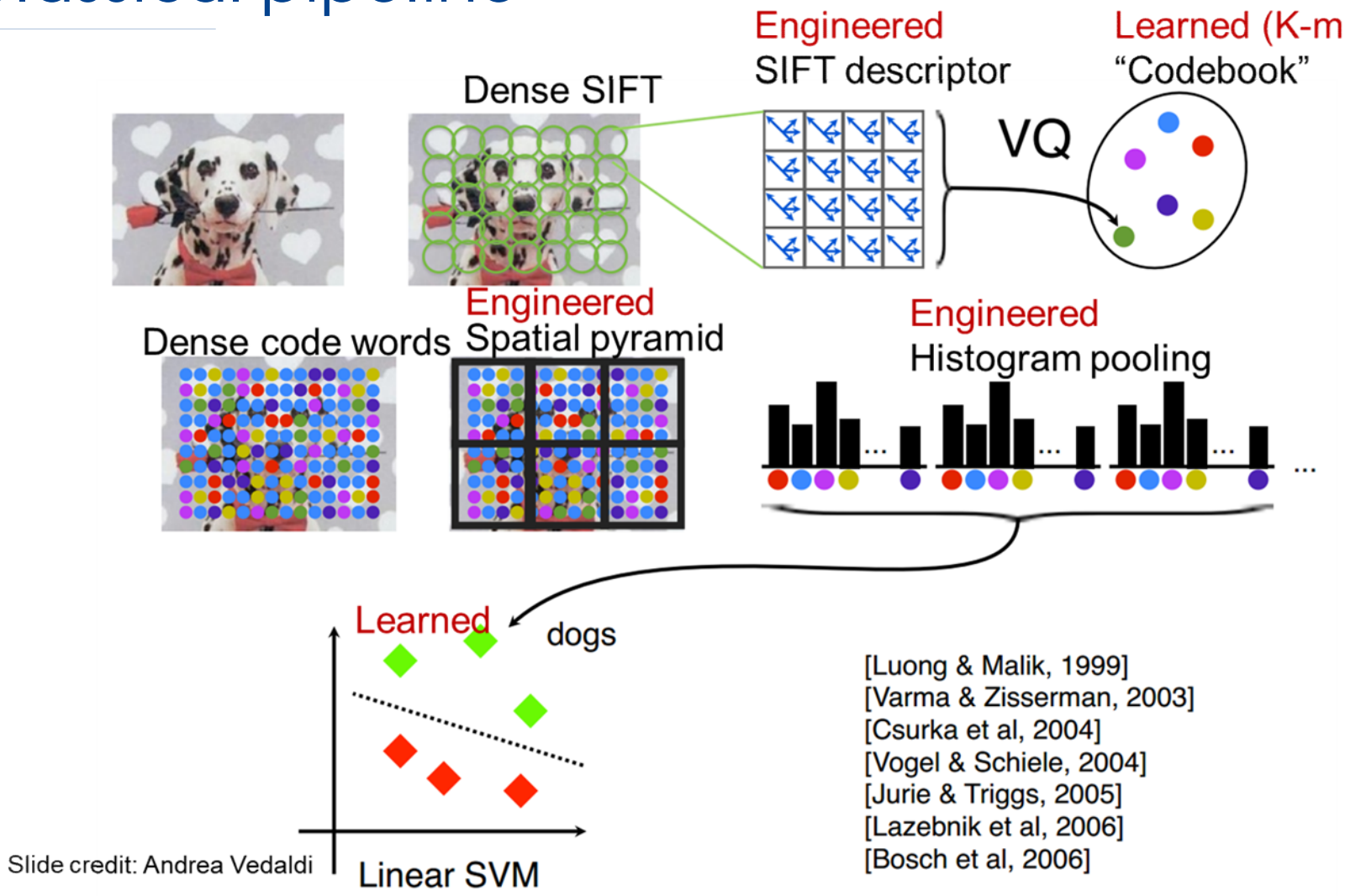
Tra…able Classifier

"Classic car"

End-to-end learning

Classical pipeline

In-between manual feature extraction and CNNs



Slide credit: Andrea Vedaldi

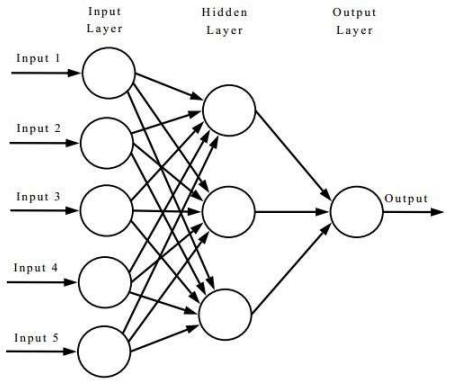
[Luong & Malik, 1999]
[Varma & Zisserman, 2003]
[Csurka et al, 2004]
[Vogel & Schiele, 2004]
[Jurie & Triggs, 2005]
[Lazebnik et al, 2006]
[Bosch et al, 2006]

Current approach

Architecture

› Build networks with many interconnected neurons and many layers (Deep Learning)

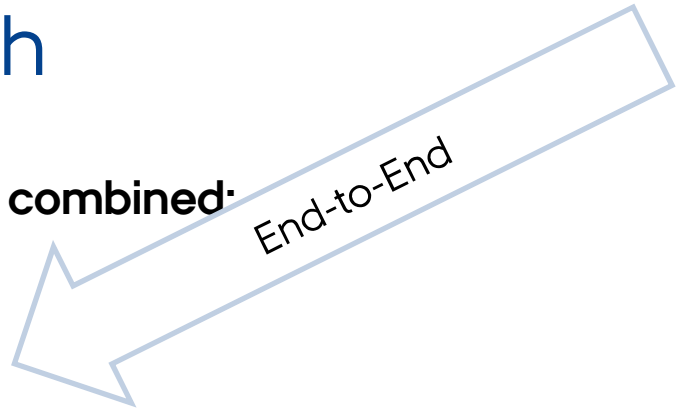
No theory about Deep Learning and how to construct the architecture of the NN. Lots of research is done by trail and error.



Training

› Optimization problem (highly non-convex)

Current approach



Artificial Neural Networks for combined:

- Image representation
- Classification

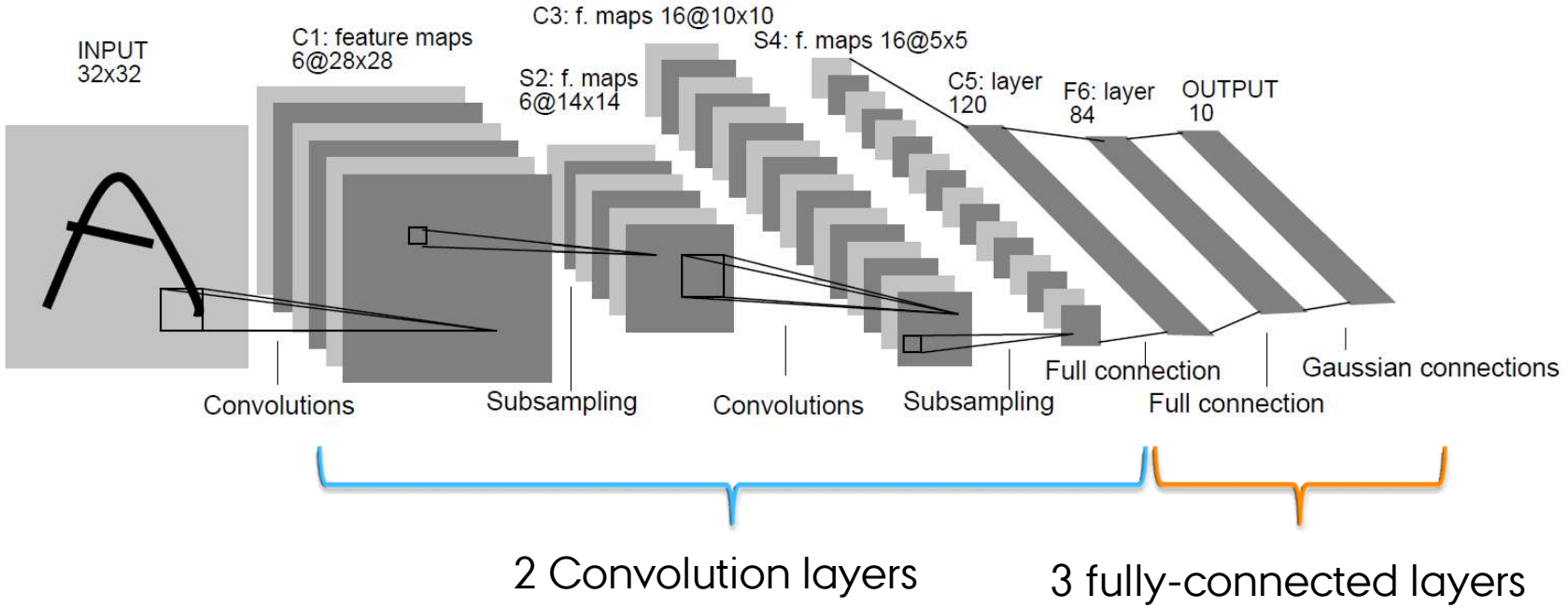
Representation learning:

- Image convolution using 2D (or 3D) filters
- Local region pooling

Classification:

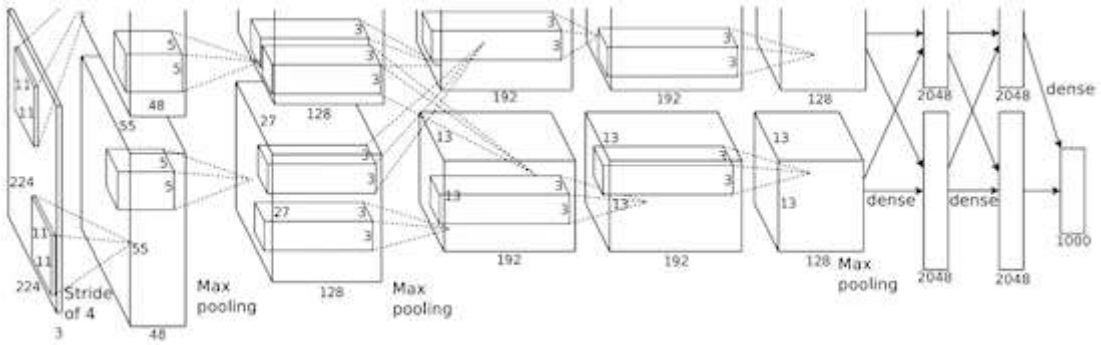
- Fully connected layers in Front-End of CNN's

CNN architecture



LeCun Yann, et .al , "Gradient-Based Learning Applied to Document Recognition" 1998.

Another CNN architecture



Alexnet consists of five convolutional layers and three fully-connected layers. The classifier is softmax.

The Alexnet (2012) achieved the best results on ImageNet, decreasing the state-of-the-art error rate from 47.1% to 37.5%.

Krizhevsky Alex, "ImageNet Classification with Deep Convolutional", Neural Networks" 2012.

Another CNN architecture

The knowledge of the Alexnet networks is stored in the parameters.
Alexnet has in total ~60.000.000 parameters to be calculated.

Trained on a subset of ImageNet with roughly 1000 images in each of 1000 categories. In all, there are roughly 1.2 million training images, 50,000 validation images, and 150,000 testing images.

Success from

- the efficient use of GPUs,
- ReLUs faster convergence using ReLU
- new regularization technique called dropout, and techniques to generate more training examples by deforming the existing ones.

Krizhevsky Alex, "ImageNet Classification with Deep Convolutional", Neural Networks" 2012.

Weights as image filters (FIR)

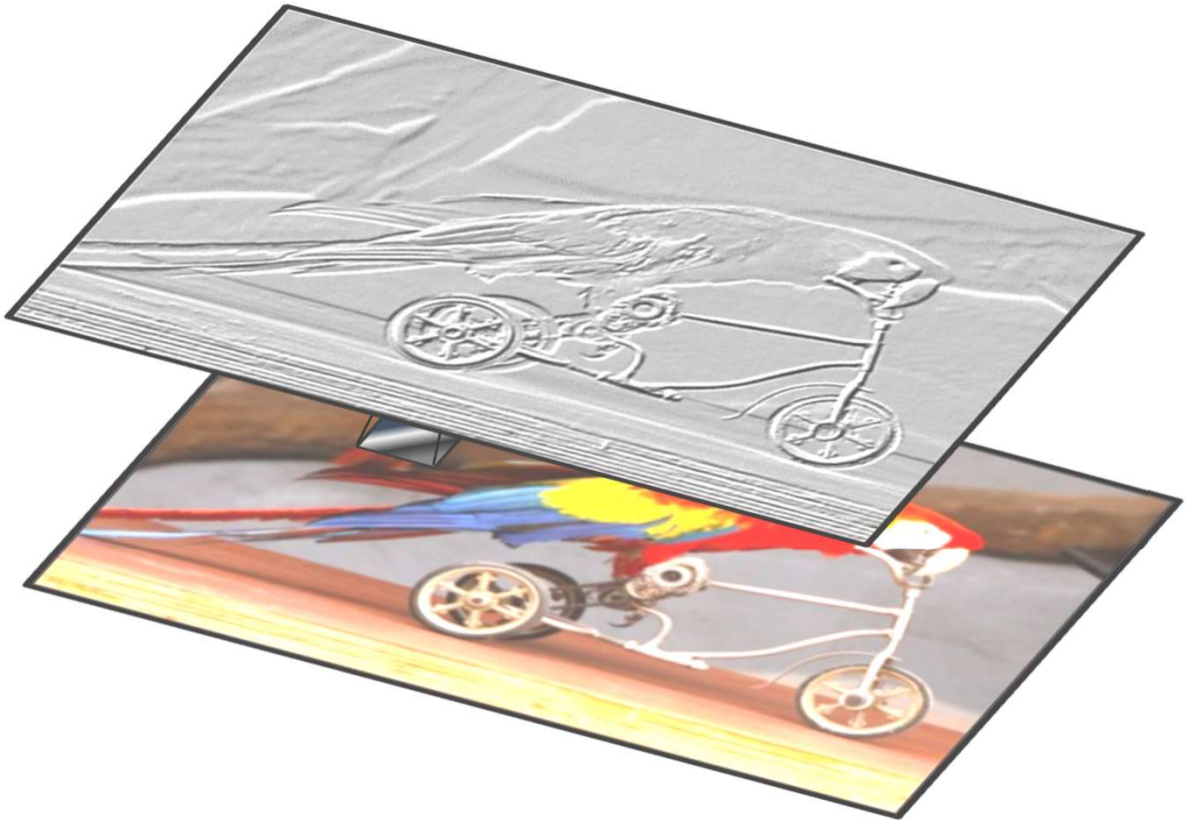


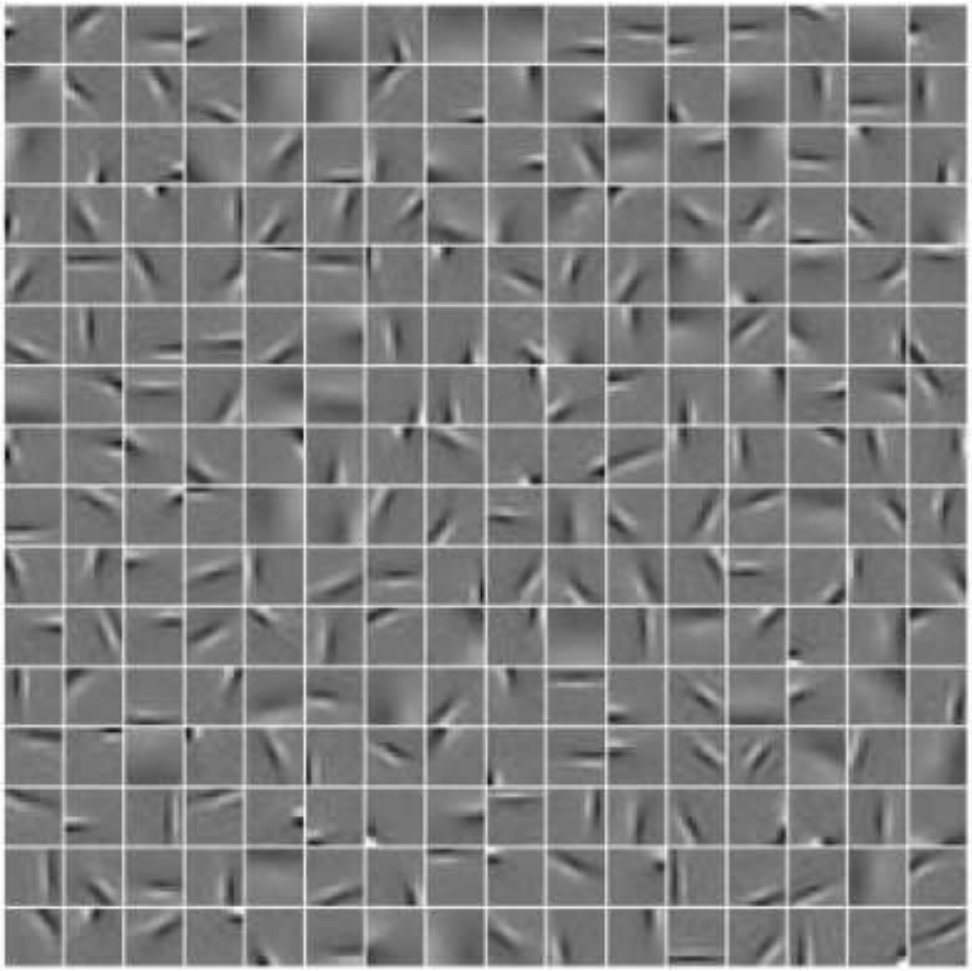
Image Representation type II

Convolution

Image Representation type I

figure credit: Larry Zitnick

Weights as image filters (FIR)



Convolutional layer

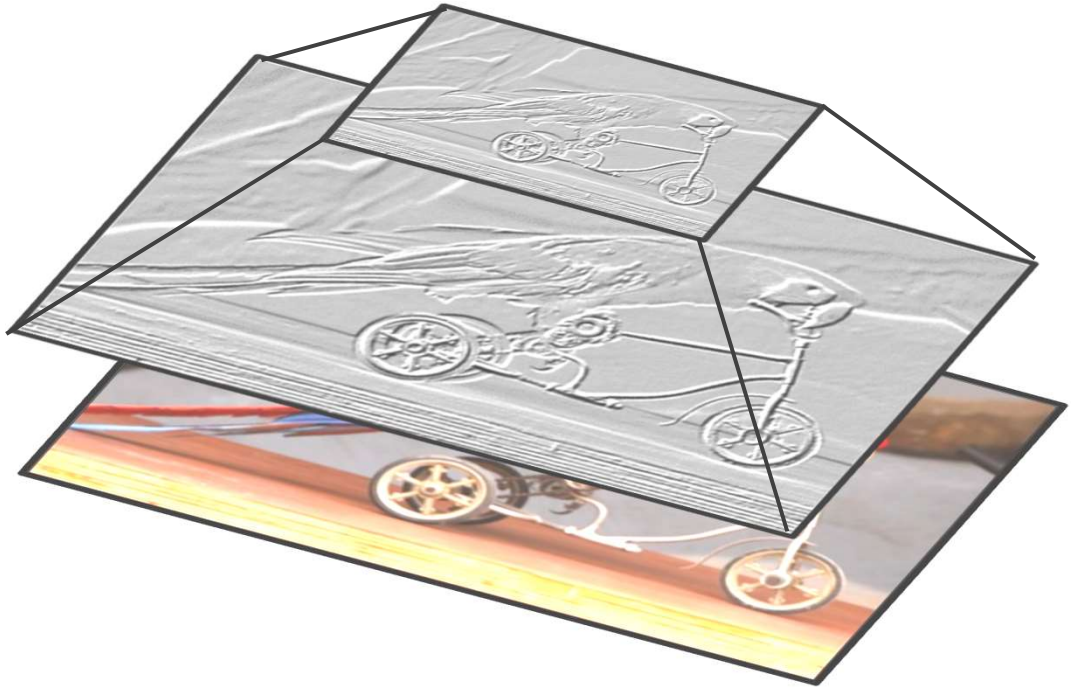


Image Representation type III

Pooling

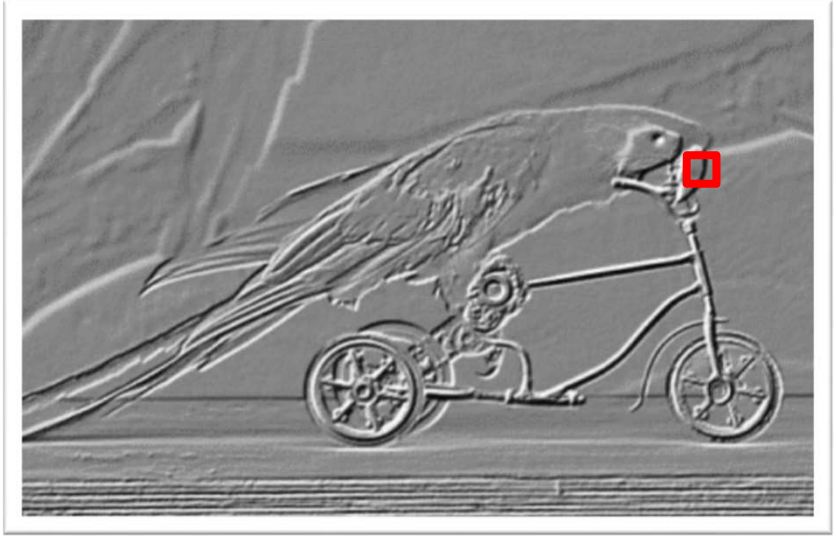
Image Representation type II

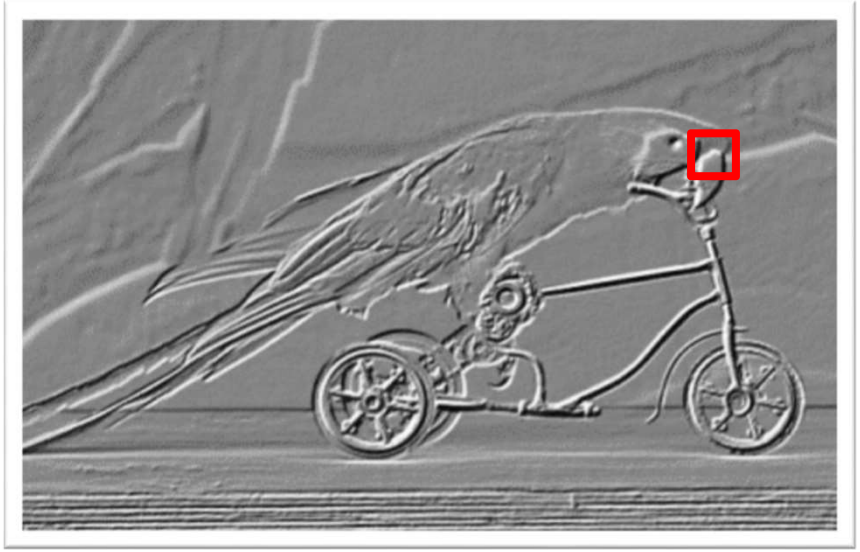
Convolution

Image Representation type I

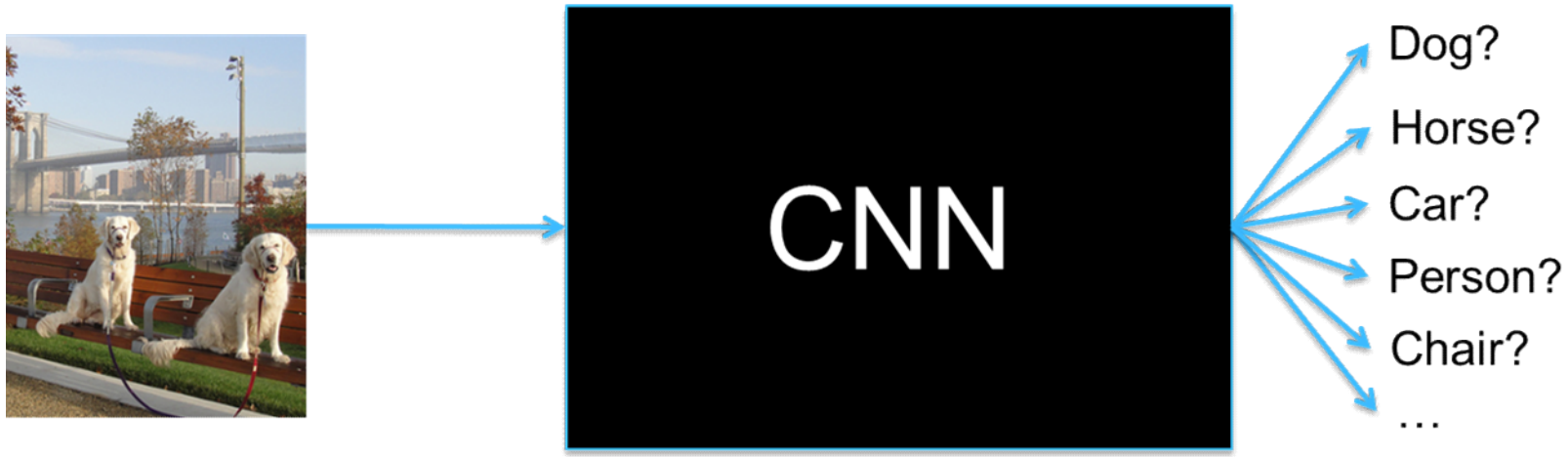
figure credit: Larry Zitnick

Weights as filters in different scales

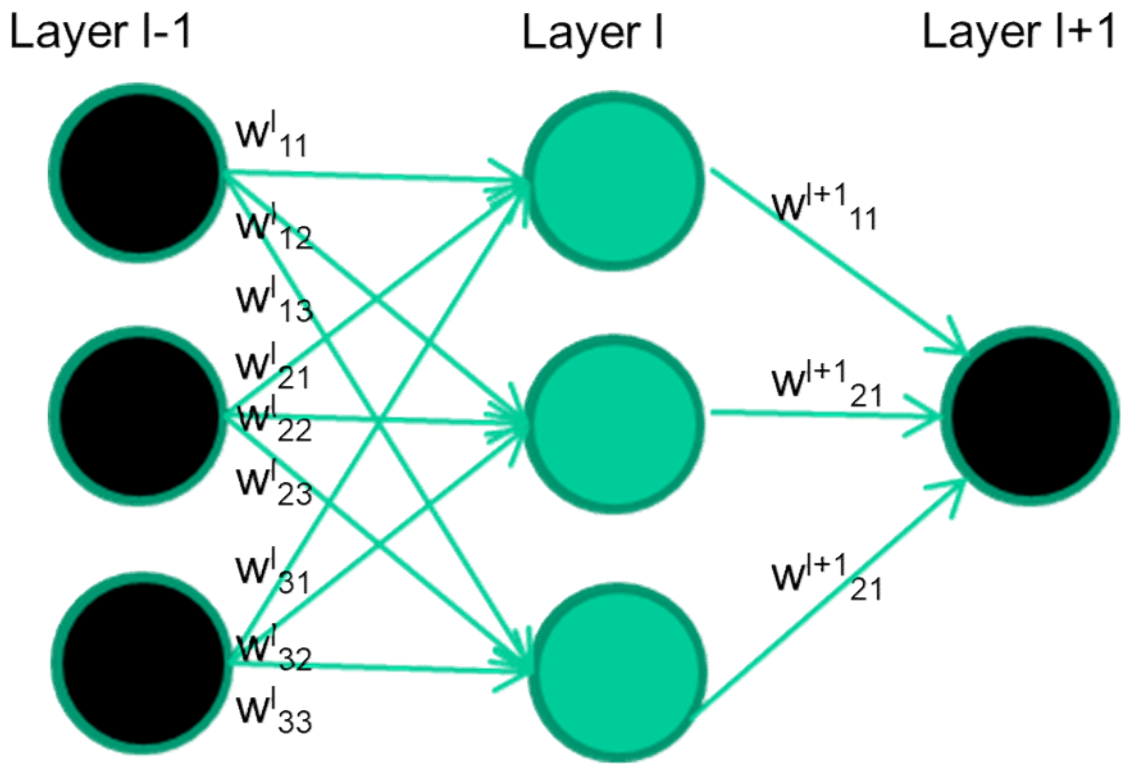




CNN training



Fully-connected layers



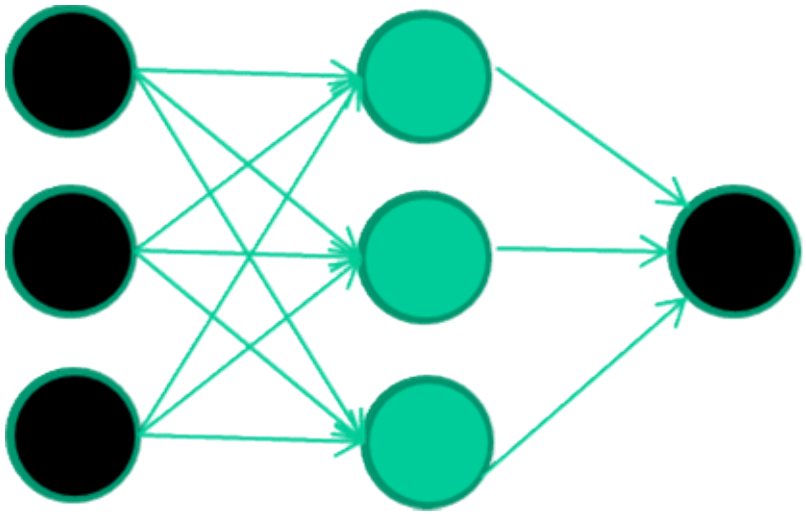
CNN training

Initialise with random weights

Training images

representations			***class***
1.4	2.7	1.9	0
3.8	3.4	3.2	0
6.4	2.8	1.7	1
4.1	0.1	0.2	0

etc …

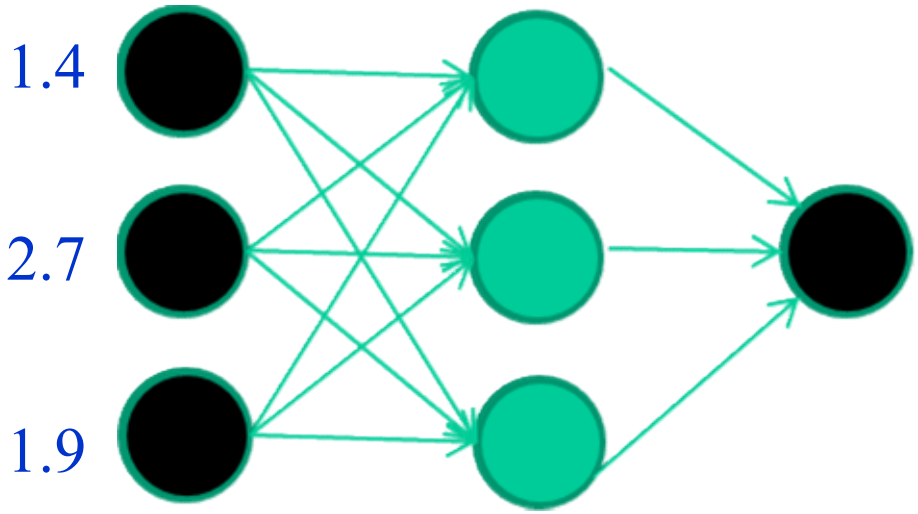


CNN training

Training images

representations			***class***
1.4	2.7	1.9	0
3.8	3.4	3.2	0
6.4	2.8	1.7	1
4.1	0.1	0.2	0

etc …

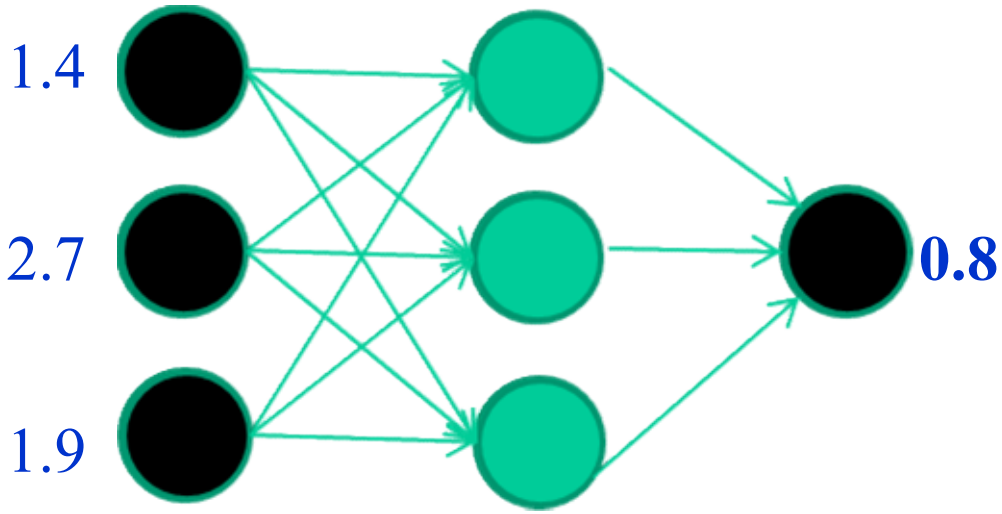


CNN training

Training images

representations			***class***
1.4	2.7	1.9	0
3.8	3.4	3.2	0
6.4	2.8	1.7	1
4.1	0.1	0.2	0

etc …



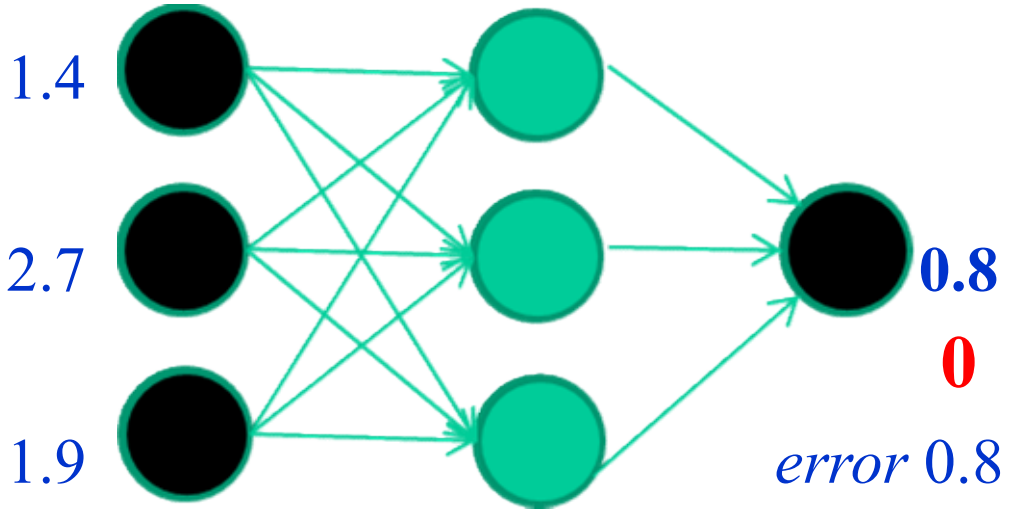
Network weights adaptation

CNN training

Training images

representations			***class***
1.4	2.7	1.9	0
3.8	3.4	3.2	0
6.4	2.8	1.7	1
4.1	0.1	0.2	0

etc …

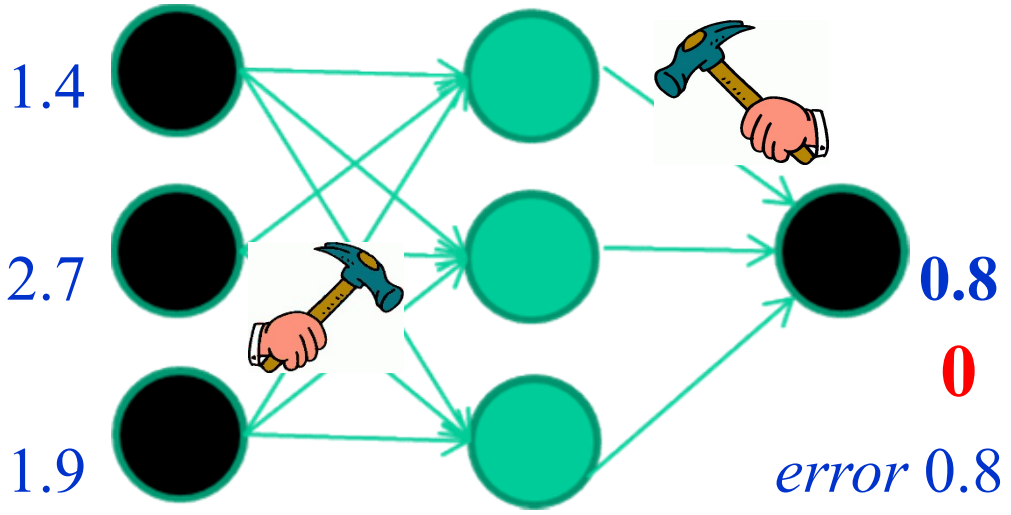


CNN training

Training images

representations			***class***
1.4	2.7	1.9	0
3.8	3.4	3.2	0
6.4	2.8	1.7	1
4.1	0.1	0.2	0

etc …

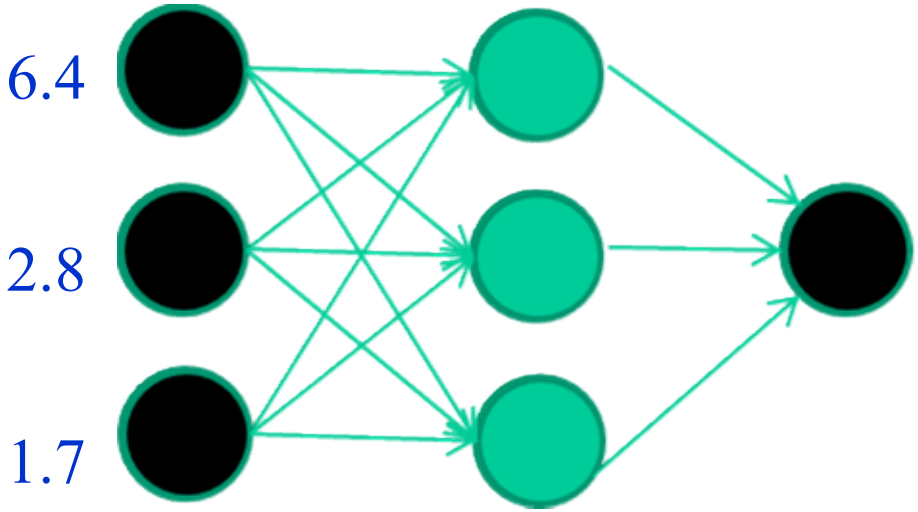


CNN training

Training images

representations			***class***
1.4	2.7	1.9	0
3.8	3.4	3.2	0
6.4	2.8	1.7	1
4.1	0.1	0.2	0

etc …

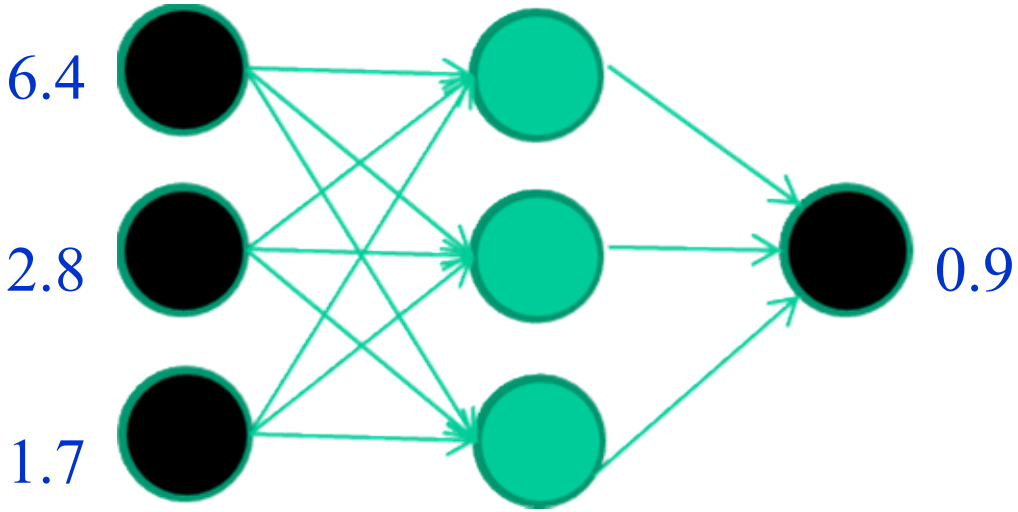


CNN training

Training images

representations			***class***
1.4	2.7	1.9	0
3.8	3.4	3.2	0
6.4	2.8	1.7	1
4.1	0.1	0.2	0

etc …

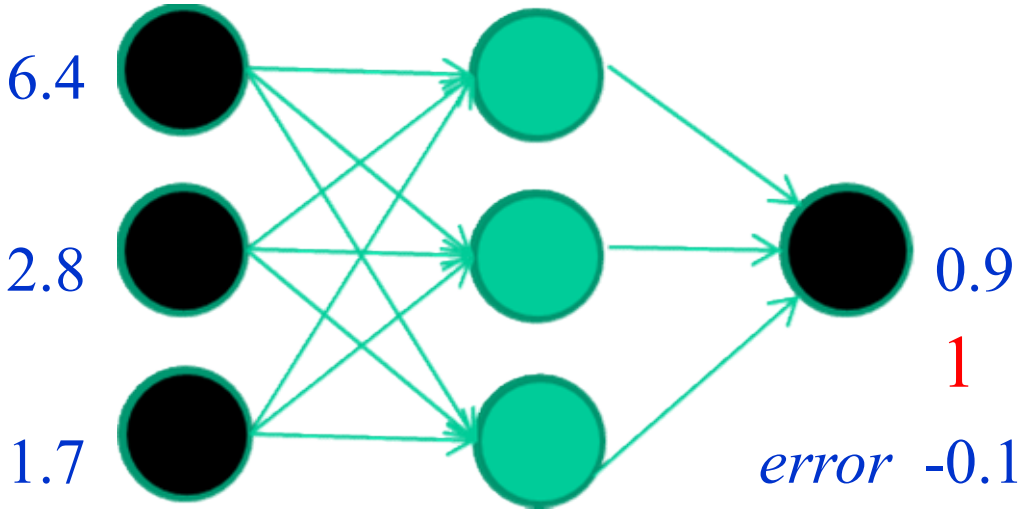


CNN training

Training images

representations			***class***
1.4	2.7	1.9	0
3.8	3.4	3.2	0
6.4	2.8	1.7	1
4.1	0.1	0.2	0

etc …

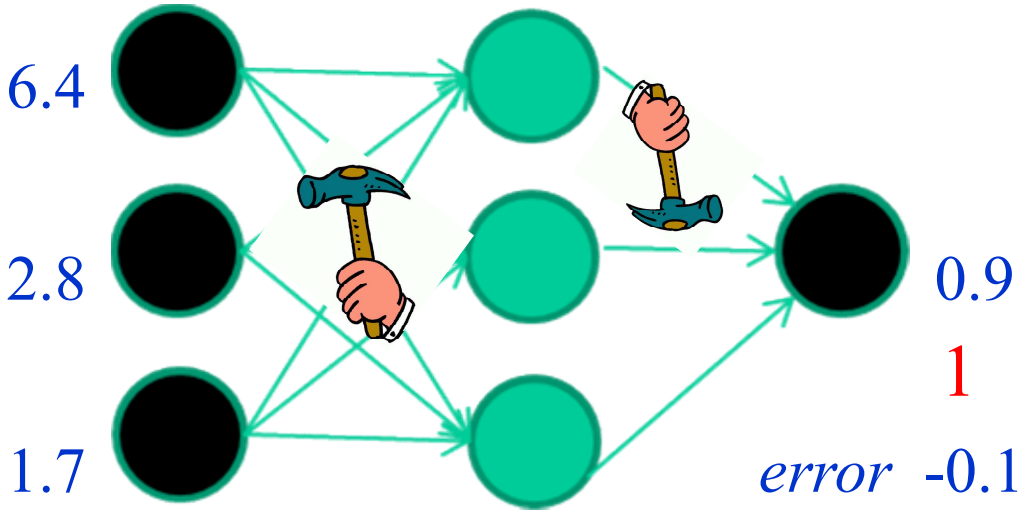


CNN training

Training images

representations			***class***
1.4	2.7	1.9	0
3.8	3.4	3.2	0
6.4	2.8	1.7	1
4.1	0.1	0.2	0

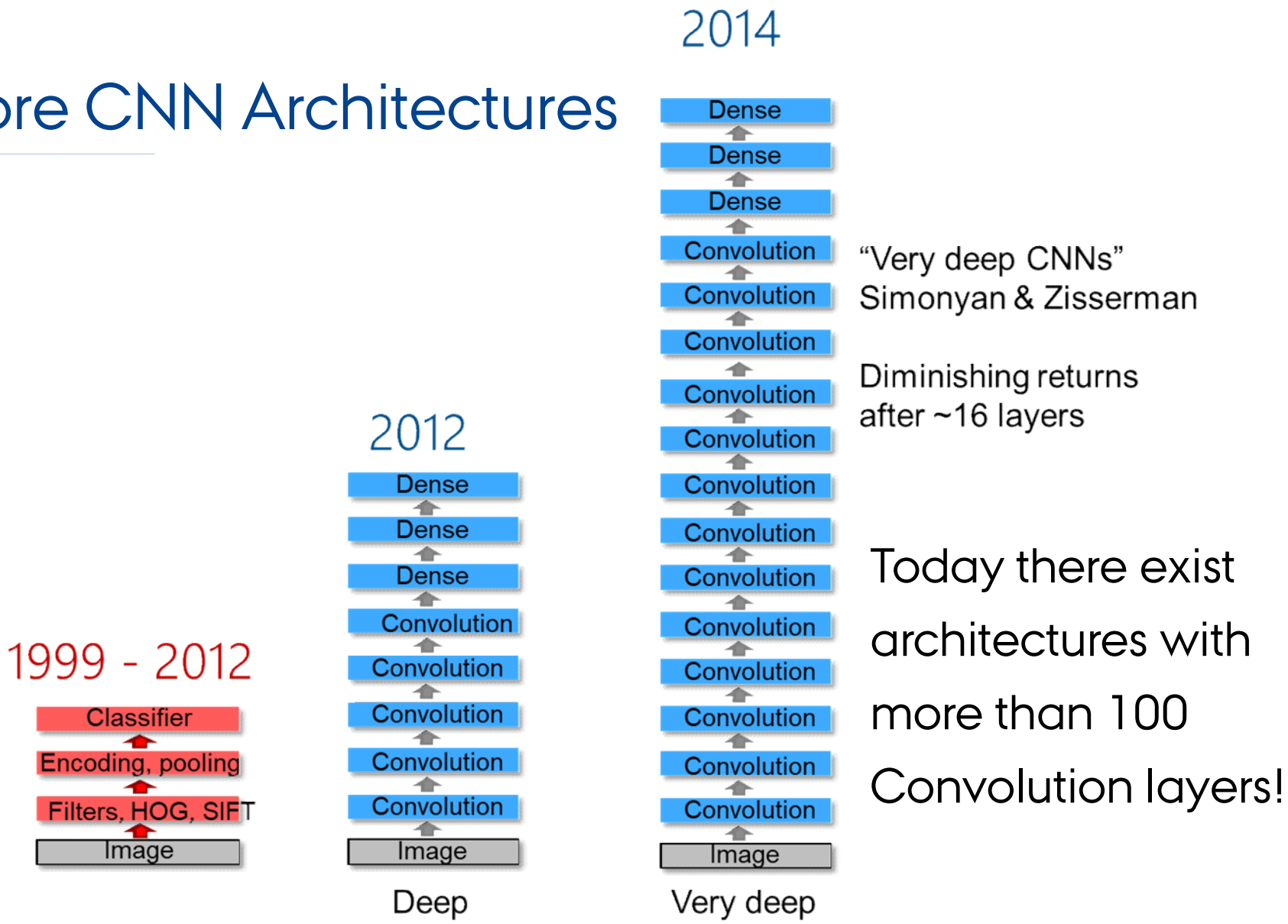
etc …



Repeat this thousands, maybe millions of times – each time taking a random training image, and making slight weight adjustments

Algorithms for weight adjustment are designed to make changes that will reduce the training error

More CNN Architectures



"Very deep CNNs"
Simonyan & Zisserman

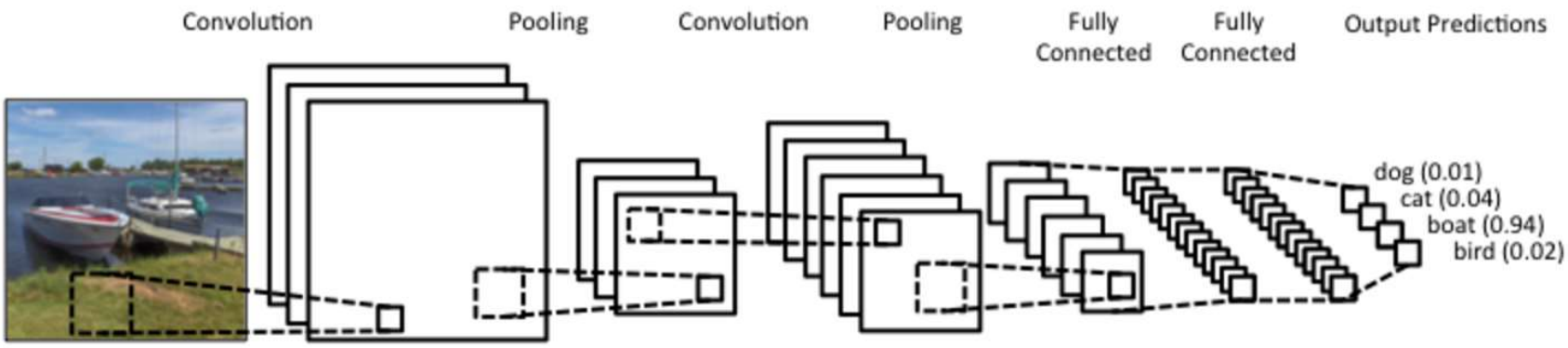
Diminishing returns
after ~16 layers

Today there exist architectures with more than 100 Convolution layers!

CNN architecture

CNN architecture:

- Convolutional layers
- Multilayer Perceptron (vector) layers



A CNN architecture

There is a vector of feature channels (e.g. RGB) at each spatial location (pixel).

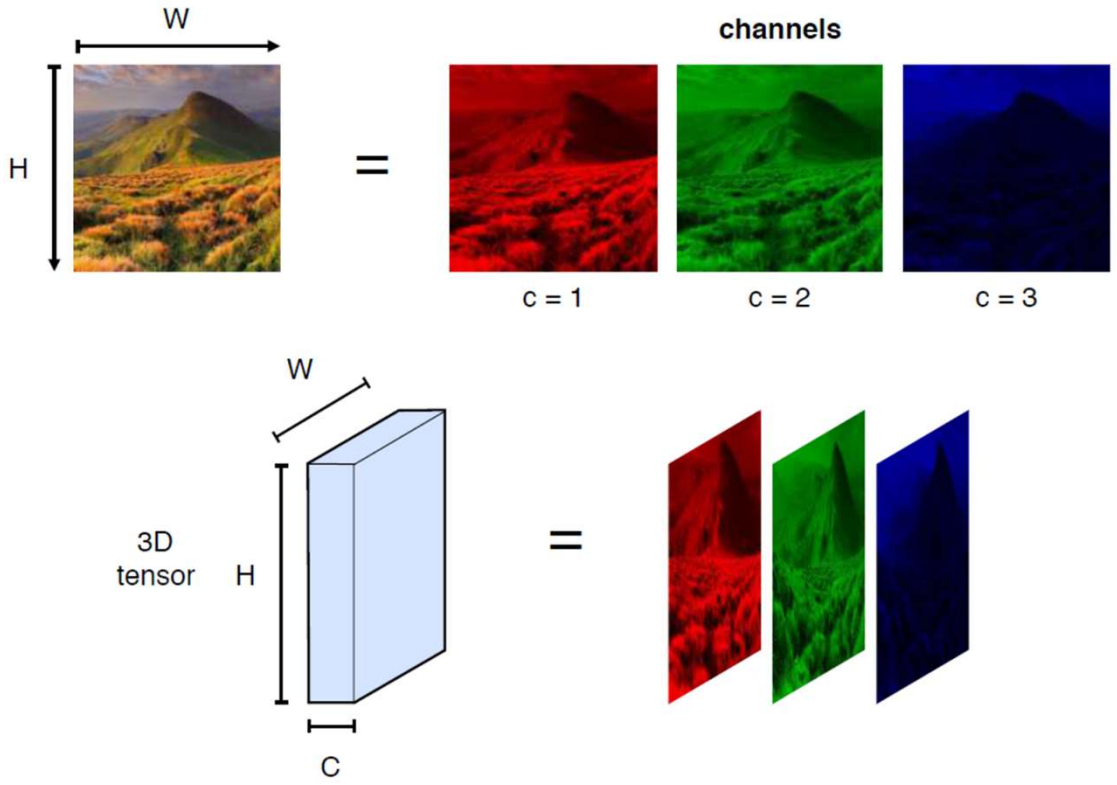


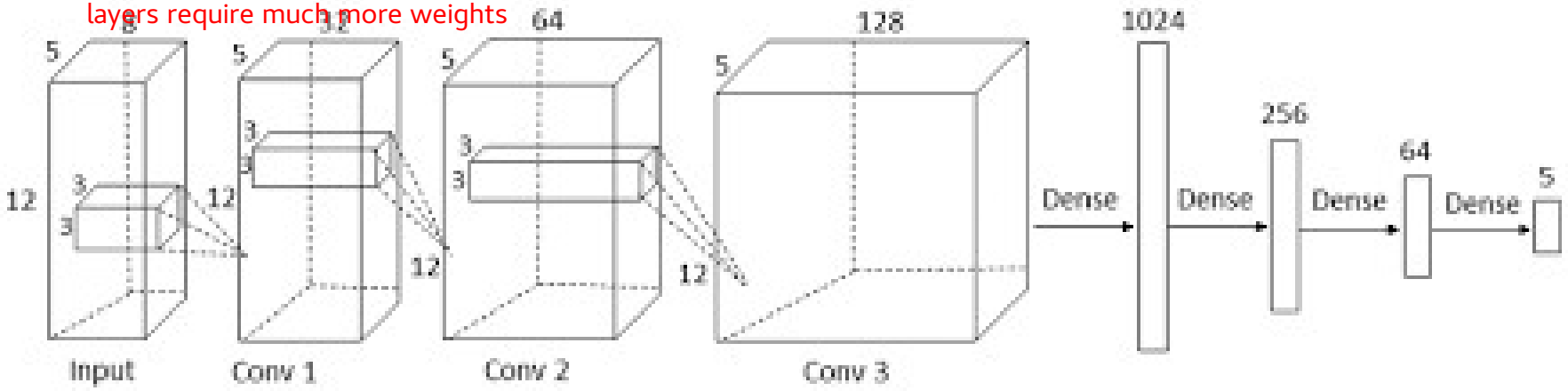
figure credit: Larry Zitnick

Convolutional Neural Networks

CNN architecture:

- Convolutional layers
- Multilayer Perceptron (vector) layers

Convolution layers have smaller number of weights because each 3x3 filter needs only 9 weights. The Fully Connected layers require much more weights



Real CNN architecture: CLs are tensors!

Object detection

Detect objects in image regions

- Sliding Bounding Box
- Pyramid-based classification for multiple scales

Object detection

Detect objects in image regions

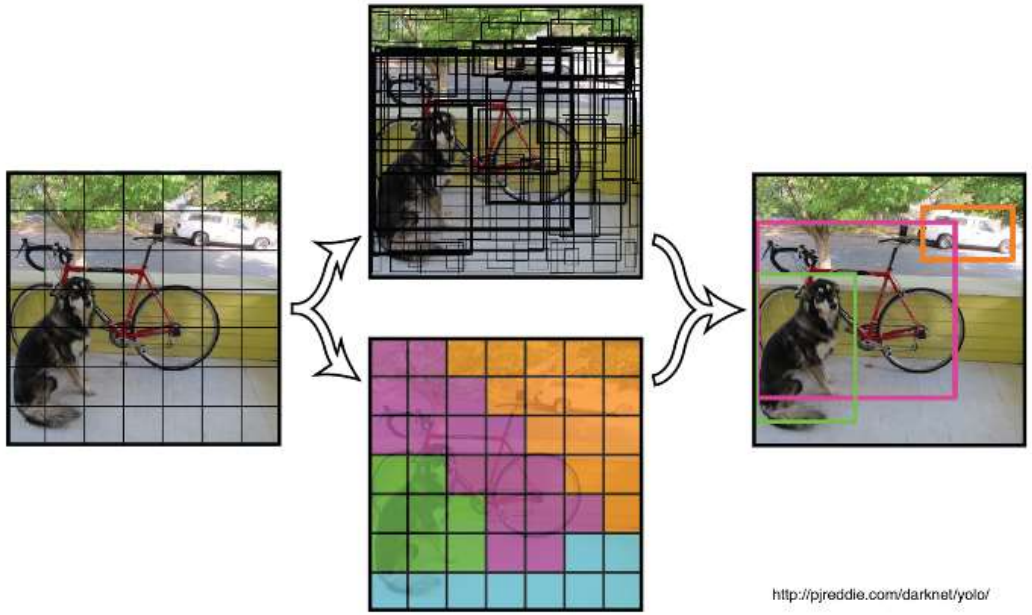
- Sliding Bounding Box
- Pyramid-based classification for multiple scales

Concurrent Object Detection and Recognition

Region proposal CNNs

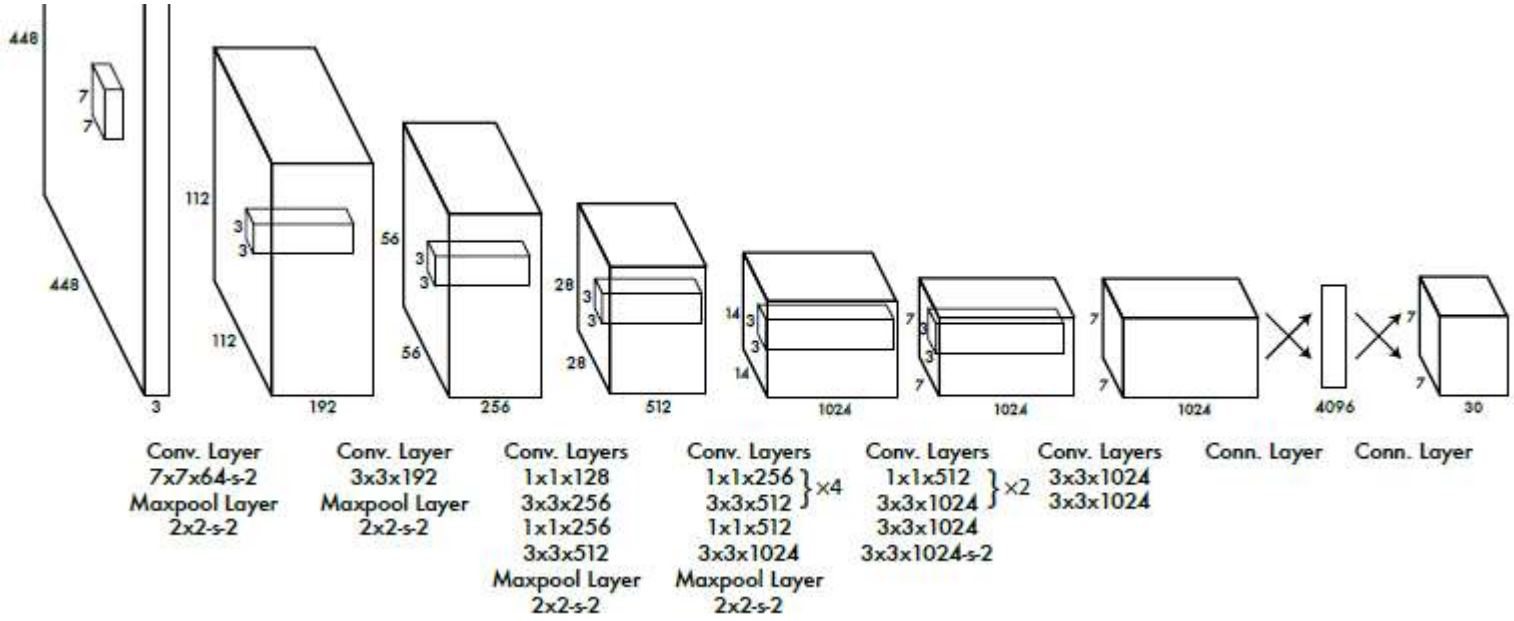
Fully-convolutional network architecture (YOLO):

- Look for one (or two) object(s) at each image sub-region
- Find the size of that object(s)
- Classify the objects



Joseph Redmon et al., "You Only Look Once: Unified, Real-Time Object Detection", arXiv 2016..

Concurrent Object Detection and Recognition



AARHUS
UNIVERSITY