Deep Learning for Visual Computing

Prof. Dr. Bernd Freisleben
Department of Mathematics & Computer Science
University of Marburg, Germany
freisleben@uni-marburg.de

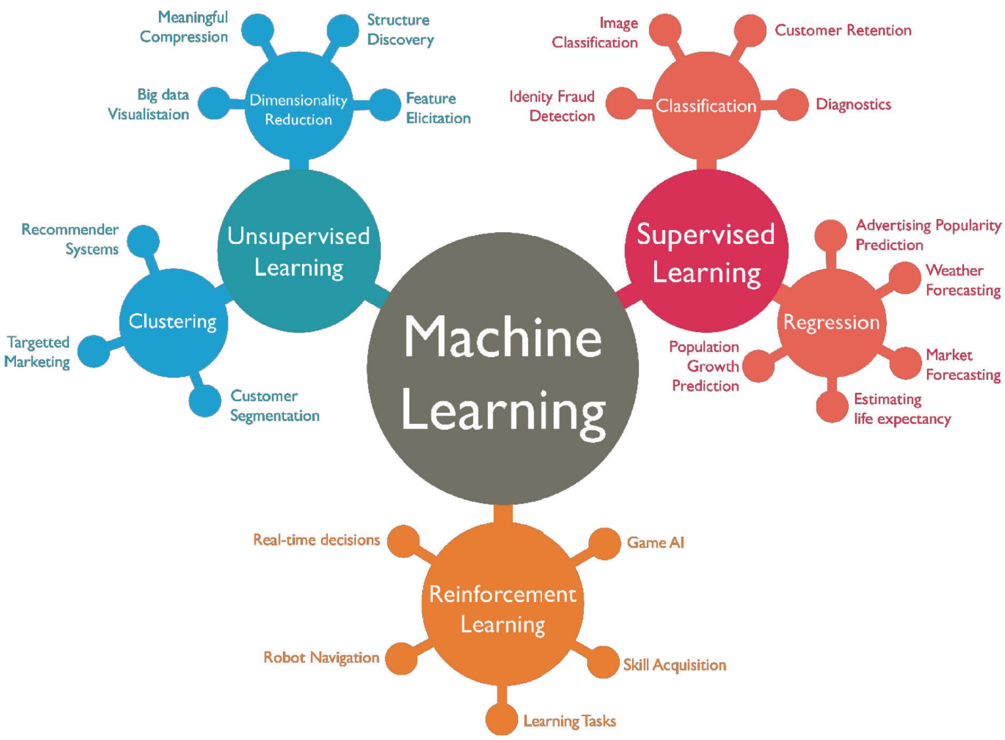
Overview

- Deep Learning
- Visual Computing
- Deep Learning for Visual Computing @ Marburg
  - Semantic Segmentation
  - Object Detection
  - Concept Detection / Person Recognition / Text Spotting
  - Similarity Search
- Conclusion

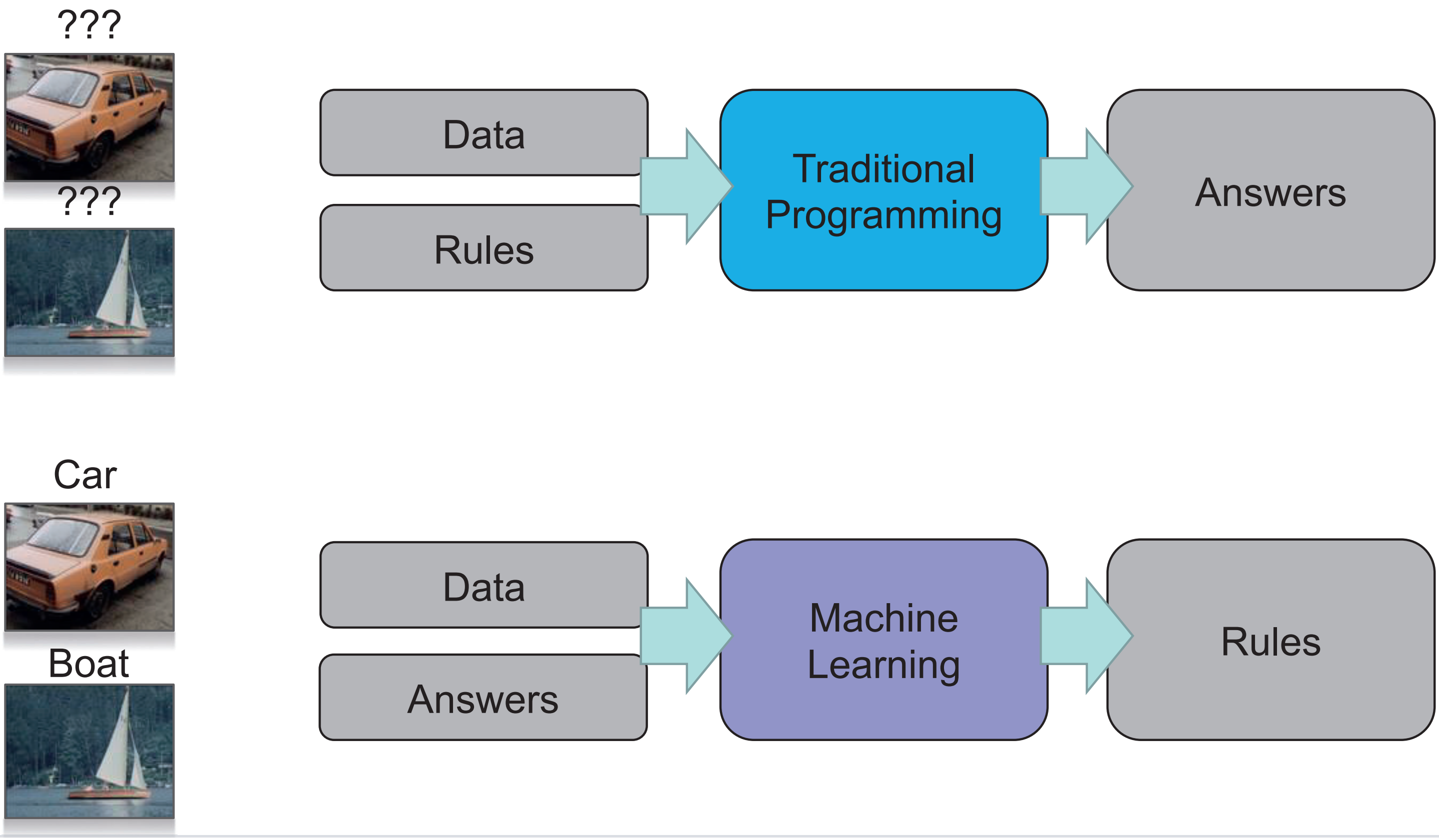
Deep Learning

Machine Learning

"Field of study that gives computers the ability to learn without being explicitly programmed"

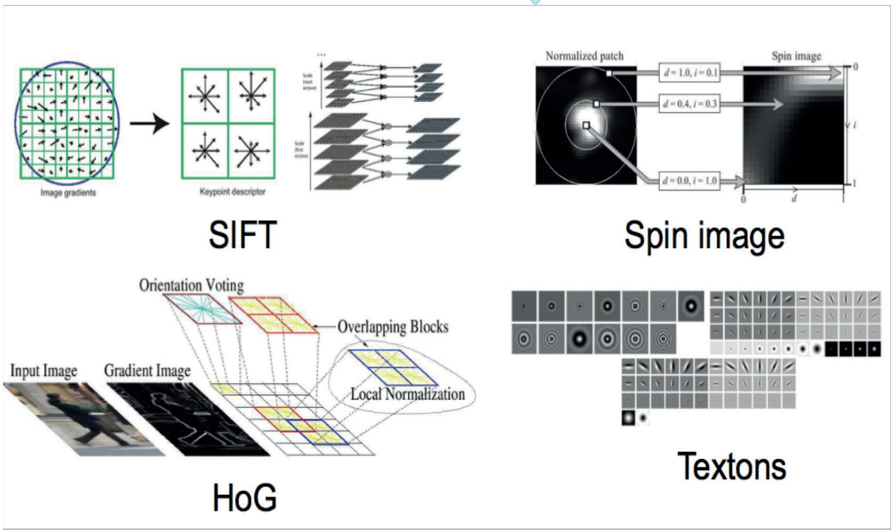


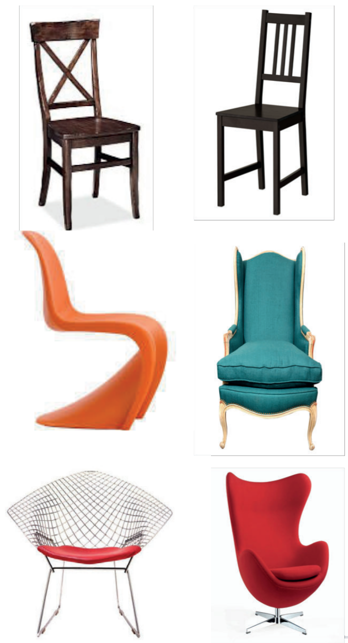
Machine Learning vs. Traditional Programming



Machine Learning for Visual Computing

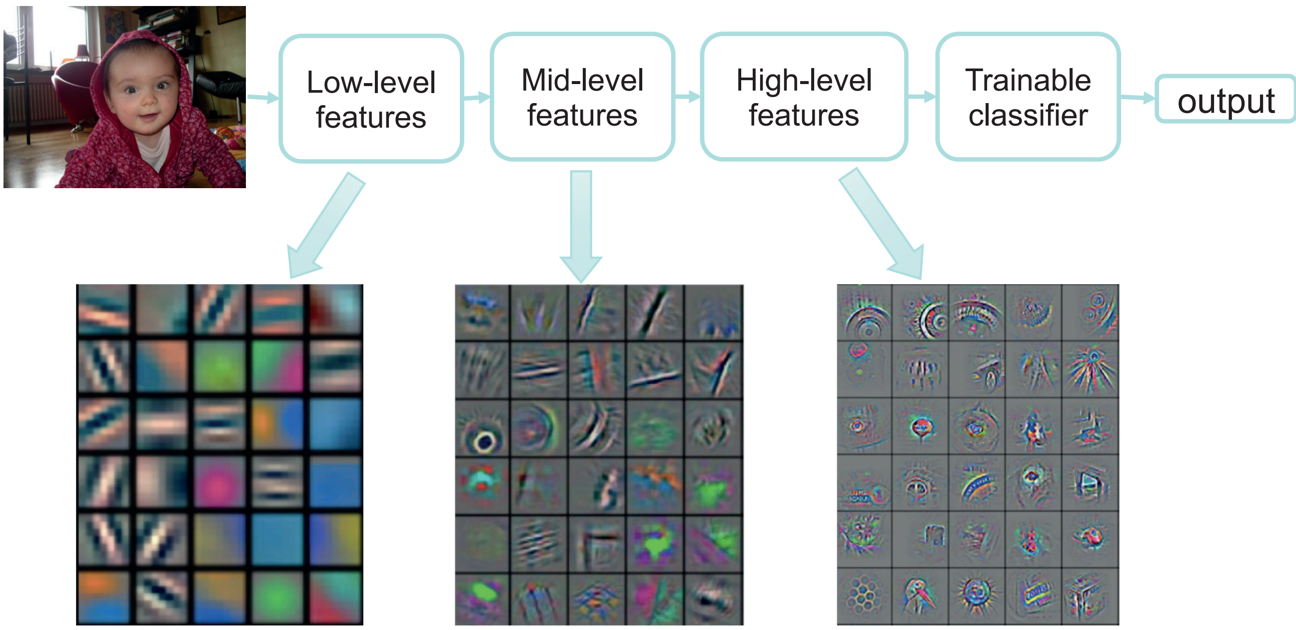
hand-crafted feature extractor → "Simple" Trainable Classifier → output





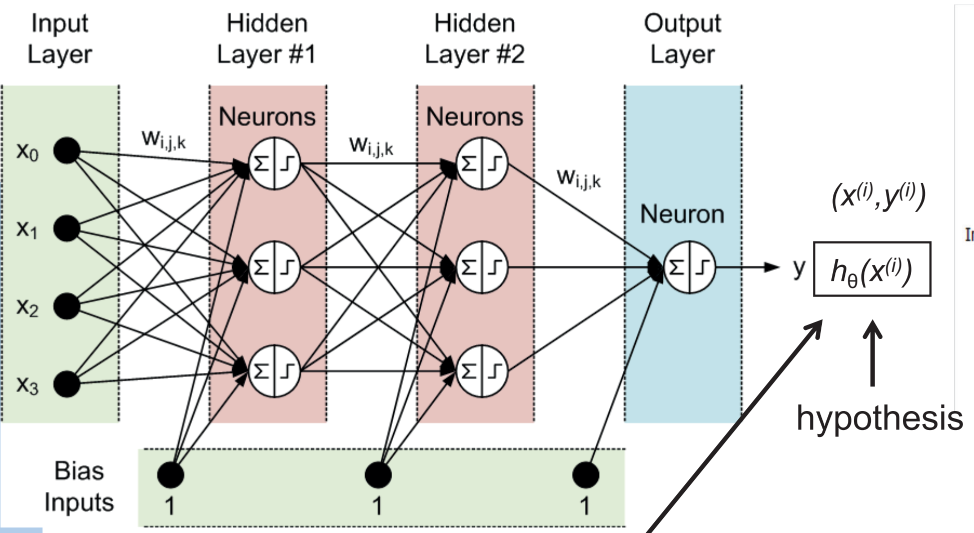
Deep Learning

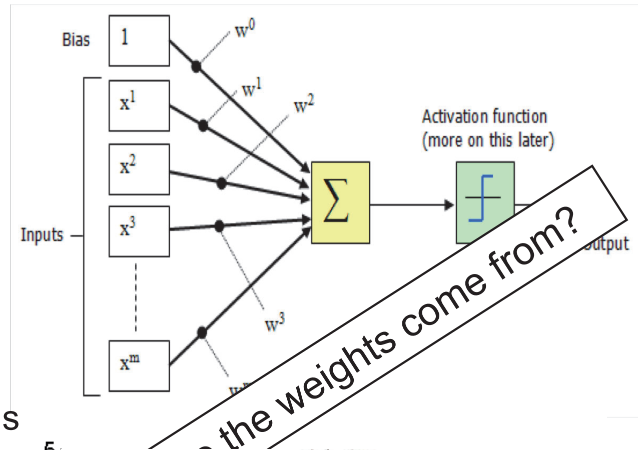
- Deep learning seeks to learn hierarchical representations (i.e., features) automatically through multiple stages of processing



Images: pixel → edge → texton → motif → part → object
Text: character → word → word group → clause → sentence → story

Deep Neural Networks

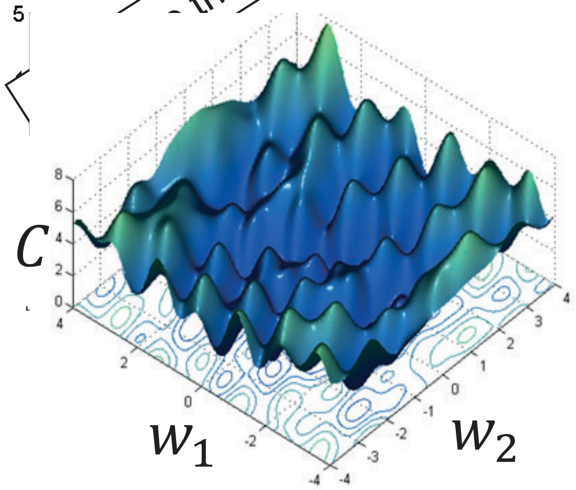




$$J(\theta_0, \theta_1, \dots, \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

Learning is adjusting the w_i,j's such that the cost function J(θ) is minimized (by gradient descent)

→ backpropagation (of errors)



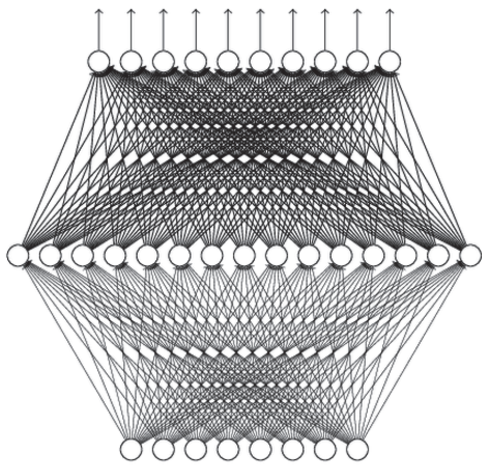
Why “deep” and not “fat”?

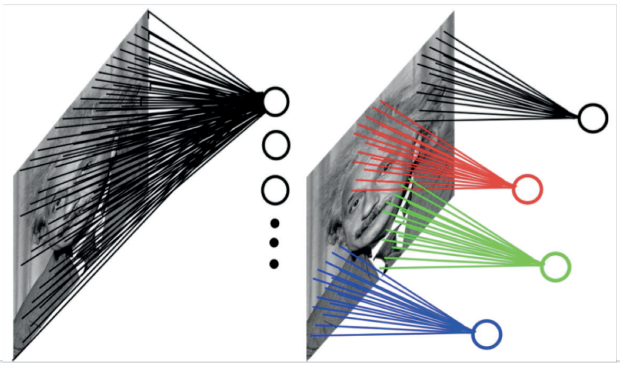
Any continuous function f

$$f : R^N \rightarrow R^M$$

can be realized by a network with one hidden layer

(given enough hidden neurons)





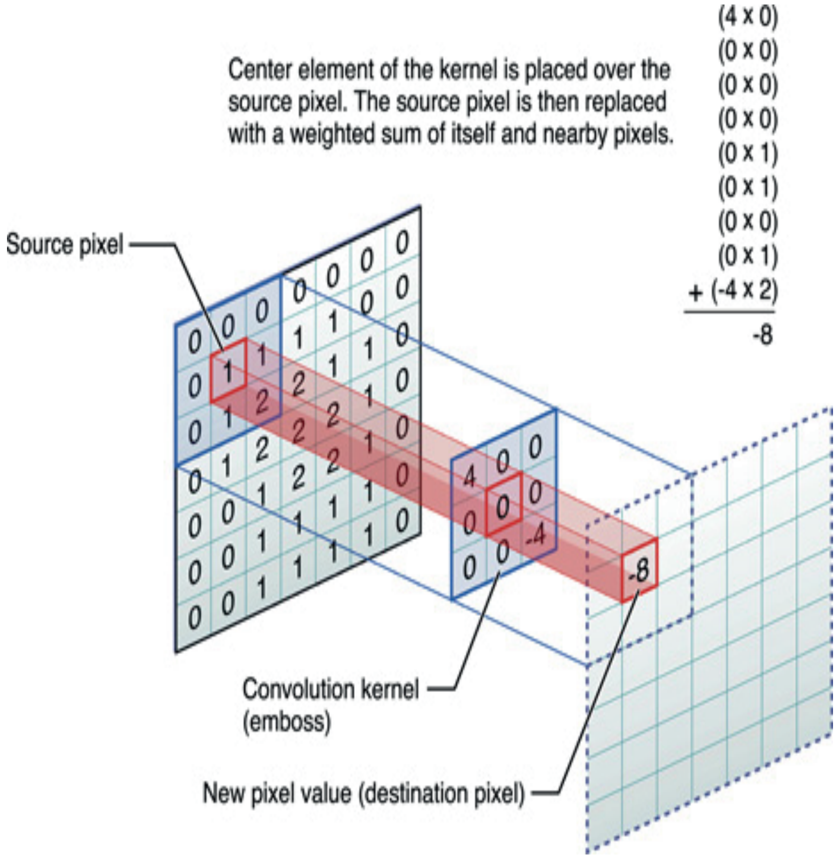
Example: 200x200 image

a) fat & fully connected: 40,000 hidden units
 => 1.6 billion parameters

a) deep & 5x5 convolution kernel: 100 feature maps
 => 2,500 parameters

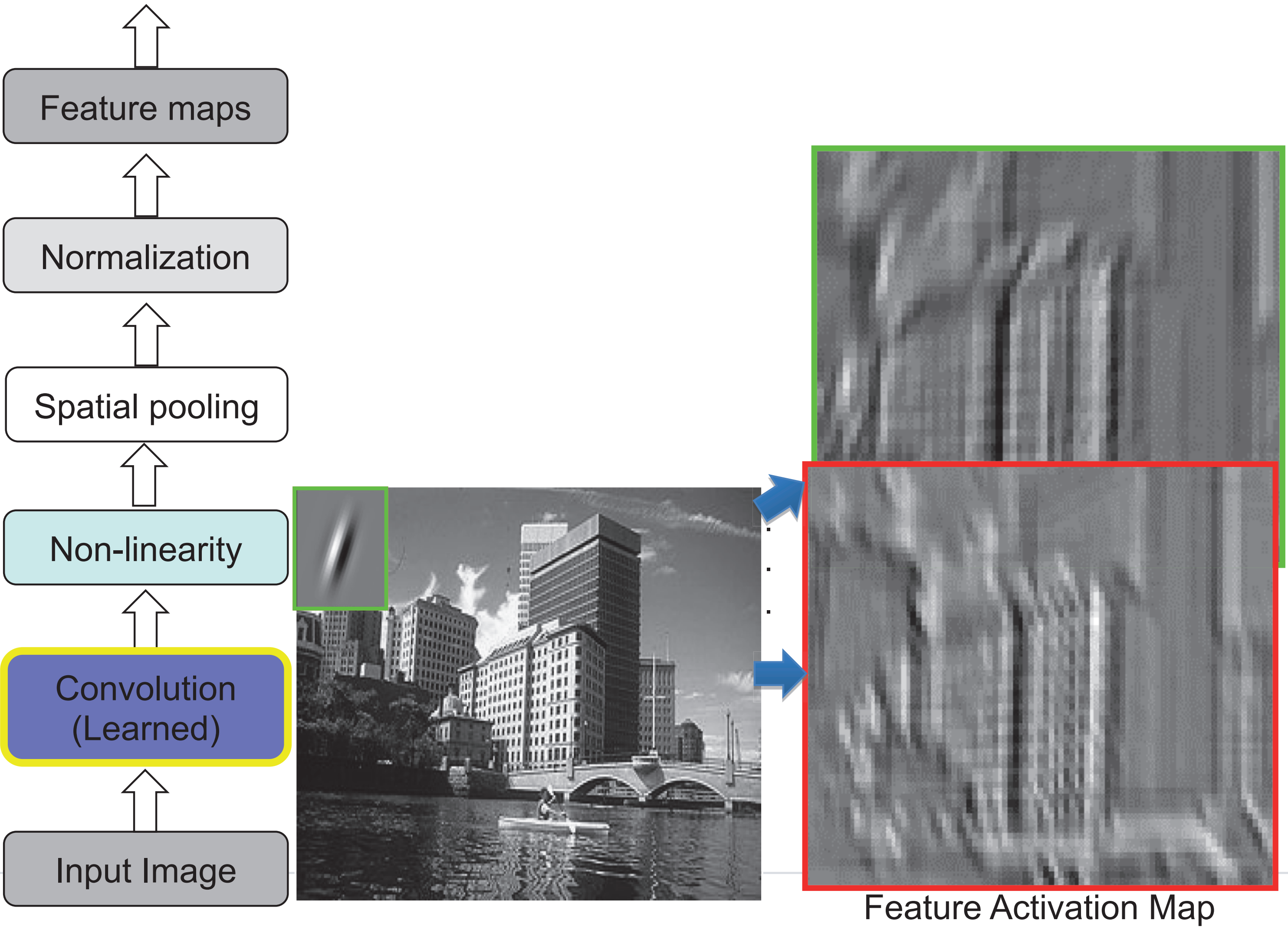
What is a Convolution?

- convolution = correlation (in image processing)
- inspired by receptive fields of the visual cortex

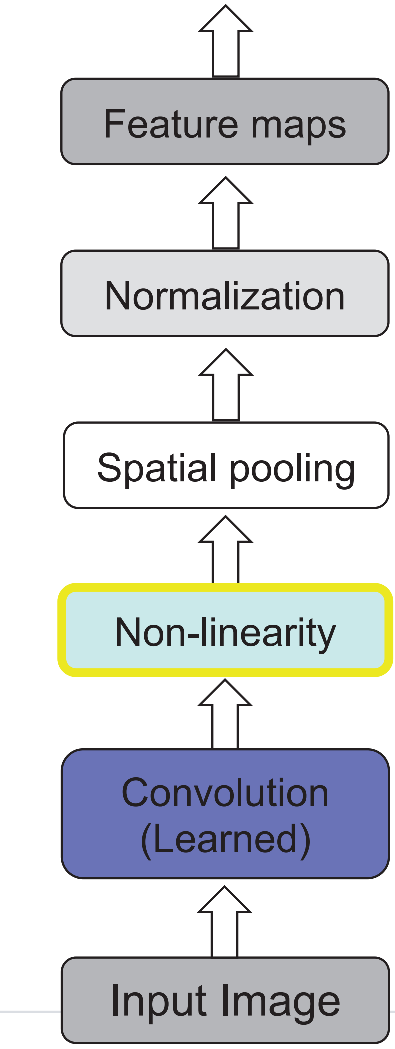


https://developer.apple.com/

Convolutional Neural Network (CNN)

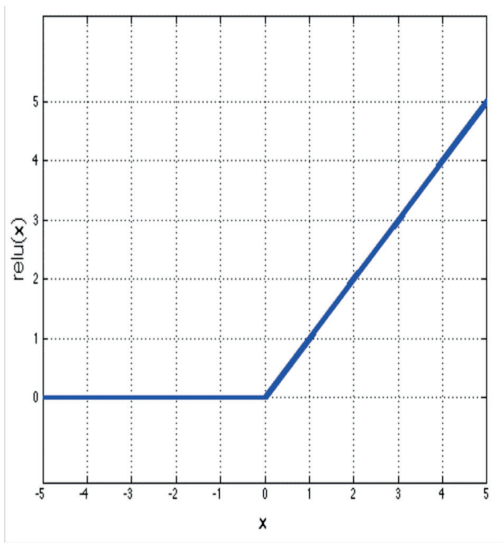


Convolutional Neural Network (CNN)

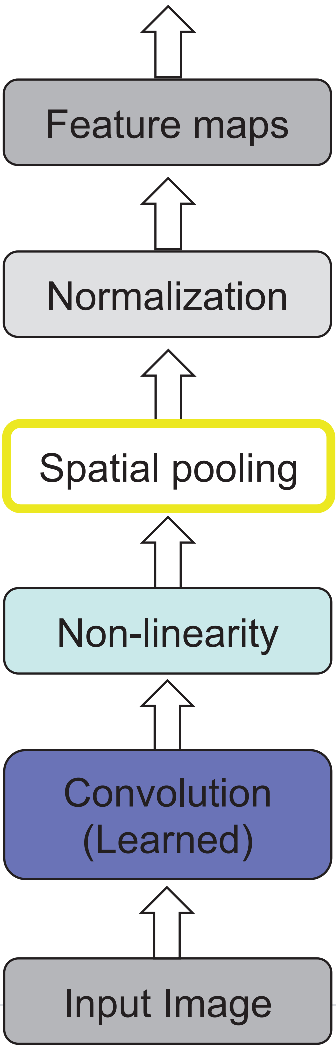


→ fast computation, no significant loss of precision

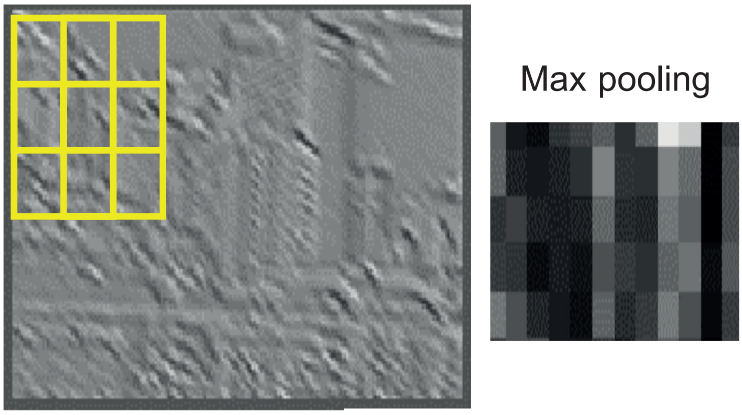
Rectified Linear Unit (ReLU)

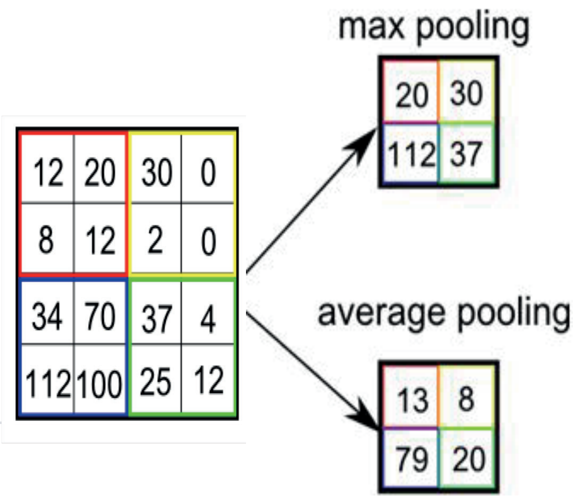


Convolutional Neural Network (CNN)

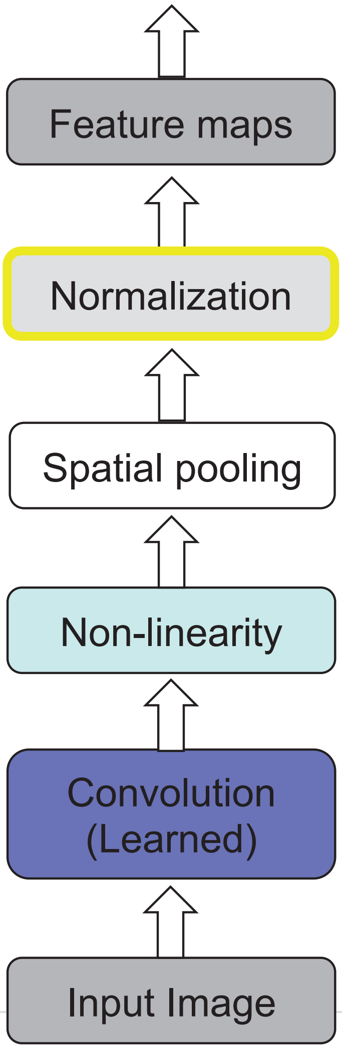


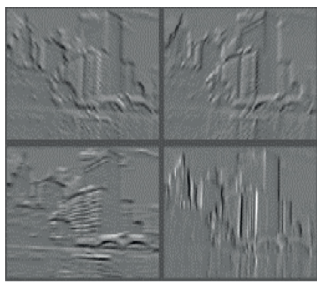
→ reduces the size of the representation
→ provides translation invariance



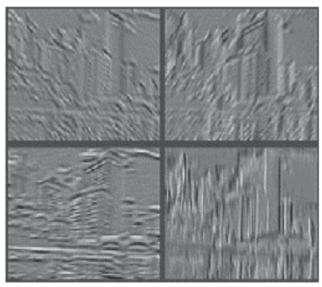


Convolutional Neural Network (CNN)





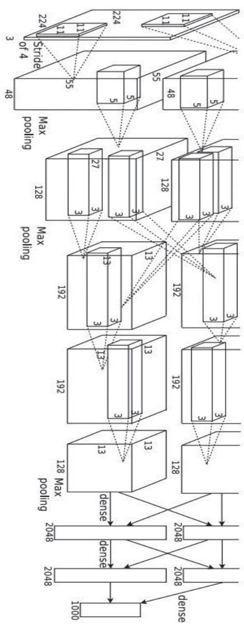
Feature Maps



Feature Maps
before
Contrast
Normalization

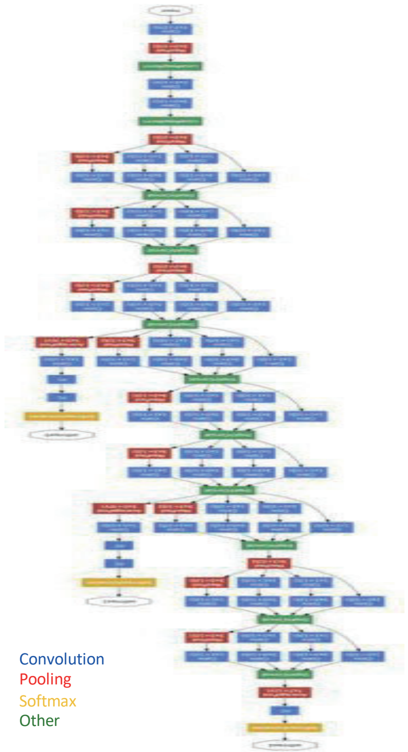
CNN Architecture Examples

AlexNet



[Krizhevsky et al. 2012]

GoogLeNet

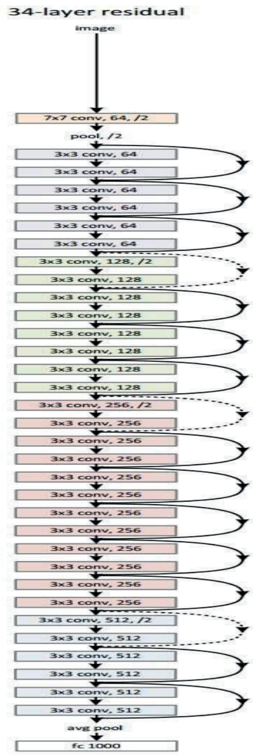


[Szegedy et al. 2014]

VGG

[Simonyan et al. 2014]

ResNet

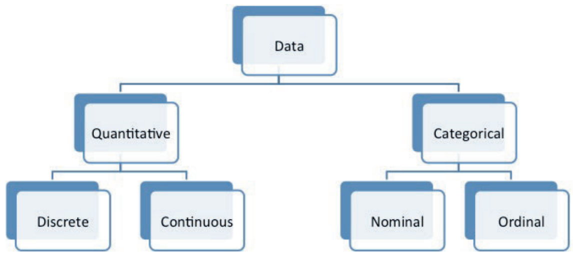


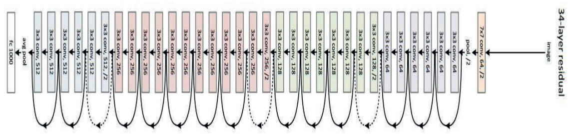
[He et al. 2015]

Recipe for Deep Learning

In theory: no need to write code!

1. Order GPU(s) + NAS
2. Install deep learning framework
3. Label data (find people)
4. Convert data (run a script)
5. Define network (edit a file)
6. Define solver (edit a file)
7. Train (pretrained weights) (run a script)





SGD, Adam, RMSprop, AdaGrad, Nesterov…

Recipe for Deep Learning: If it doesn‘t work well...

- Data preprocessing / data augmentation: check labels, mean/variance…
- Activation functions: use ReLU, try Leaky ReLU/Maxout/ELU, don’t use sigmoid…
- Weight initialization: random, pretrained, non-zero weights…
- Gradient checking: ensure backward pass is correct…
- Parameter adaptation: learning rate, momentum, batch size...
- Regularization: over/underfitting, batch normalization, dropout, weight decay...
- Architecture modification: add/remove layers, change gradient solvers...
- Evaluation: analyze/visualize internal network states, model ensembles

Deep Learning: Current Hot Topics

- Theory
  - network visualization
  - dealing with uncertainty, causal reasoning, explainable behavior
  - information bottleneck

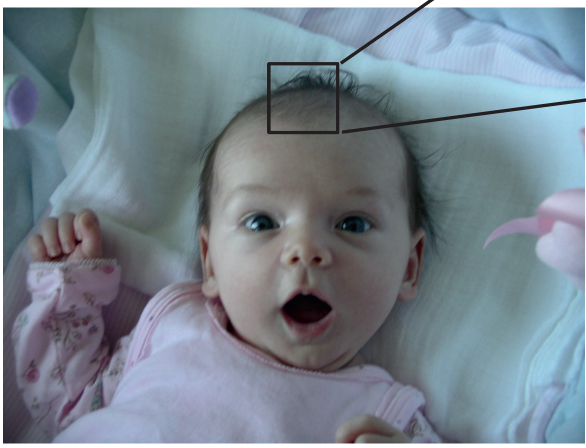
- Representation
  - data sequences
  - spatial/temporal data, data fusion
  - probabilistic relational models

- Approaches
  - cost-sensitive learning (data augmentation)
  - active learning (learning algorithm interactively queries user)
  - transfer learning (use trained algorithm for other domains)
  - ensemble learning (use multiple learning algorithms)
  - sequential learning (use recurrent neural networks)
  - semi-supervised learning (use partially labeled data)
  - unsupervised competitive learning (Generative Adversarial Networks, GANs)

Visual Computing

Problem: The Semantic Gap

What we see

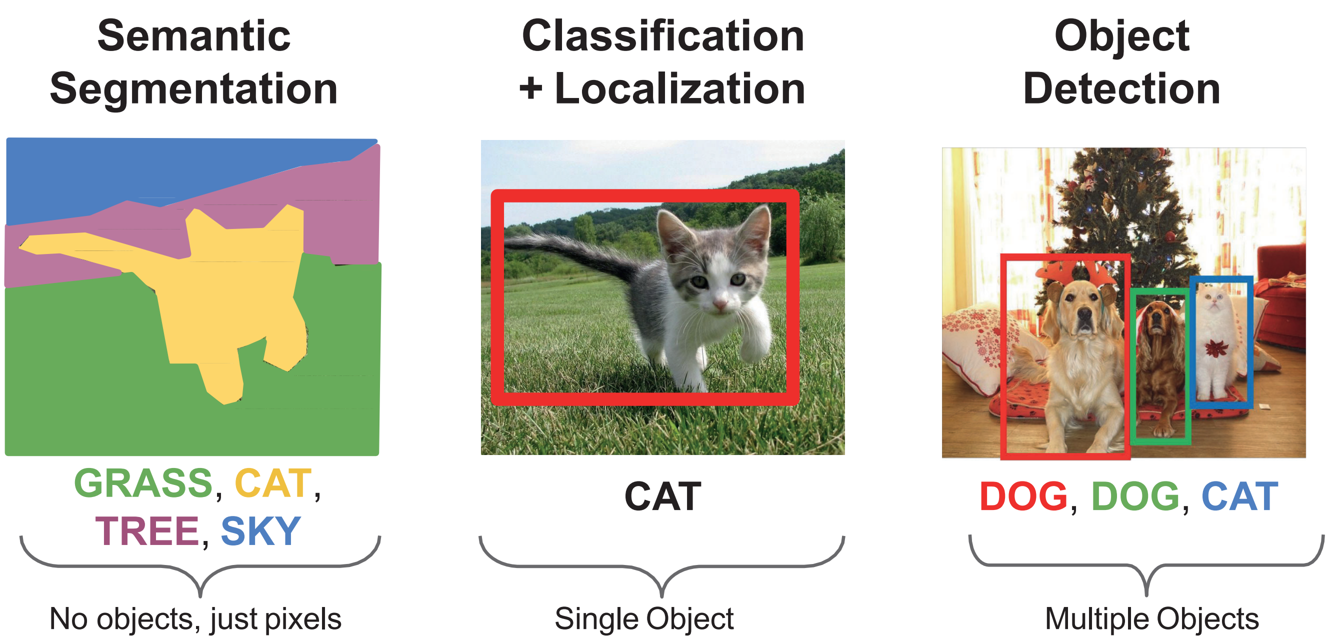


What the neural network sees

```
[[105 112 108 111 104  99 106  99  96 103 112 119 104  97  93  87]
 [ 91  98 102 106 104  79  98 103  99 105 123 136 110 105  94  85]
 [ 76  85  90 105 128 105  87  96  95  99 115 112 106 103  99  85]
 [ 99  81  81  93 120 131 127 100  95  98 102  99  96  93 101  94]
 [106  91  61  64  69  91  88  85 101 107 109  98  75  84  96  95]
 [114 108  85  55  55  69  64  54  64  87 112 129  98  74  84  91]
 [133 137 147 103  65  81  80  65  52  54  74  84 102  93  85  82]
 [128 137 144 140 109  95  86  70  62  65  63  63  60  73  86 101]
 [125 133 148 137 119 121 117  94  65  79  80  65  54  64  72  98]
 [127 125 131 147 133 127 126 131 111  96  89  75  61  64  72  84]
 [115 114 109 123 150 148 131 118 113 109 100  92  74  65  72  78]
 [ 89  93  90  97 108 147 131 118 113 114 113 109 106  95  77  80]
 [ 63  77  86  81  77  79 102 123 117 115 117 125 125 130 115  87]
 [ 62  65  82  89  78  71  80 101 124 126 119 101 107 114 131 119]
 [ 63  65  75  88  89  71  62  81 120 138 135 105  81  98 110 118]
 [ 87  65  71  87 106  95  69  45  76 130 126 107  92  94 105 112]
 [118  97  82  86 117 123 116  66  41  51  95  93  89  95 102 107]
 [164 146 112  80  82 120 124 104  76  48  45  66  88 101 102 109]
 [157 170 157 120  93  86 114 132 112  97  69  55  70  82  99  94]
 [130 128 134 161 139 100 109 118 121 134 114  87  65  53  69  86]
 [128 112  96 117 150 144 120 115 104 107 102  93  87  81  72  79]
 [123 107  96  86  83 112 153 149 122 109 104  75  80 107 112  99]
 [122 121 102  80  82  86  94 117 145 148 153 102  58  78  92 107]
 [122 164 148 103  71  56  78  83  93 103 119 139 102  61  69  84]]
```

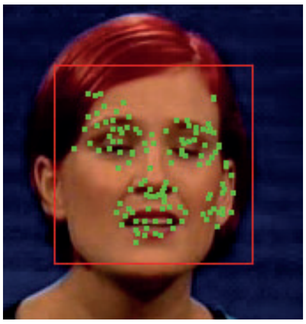
800 x 600 x 3 (3 channels RGB)

Visual Computing Tasks



Visual Computing Tasks

Person Detection + Recognition



Person X

Face, Head, Eyes, Body, Pedestrian, Crowd

Text Spotting

Gera Gewerkschafts- beratung

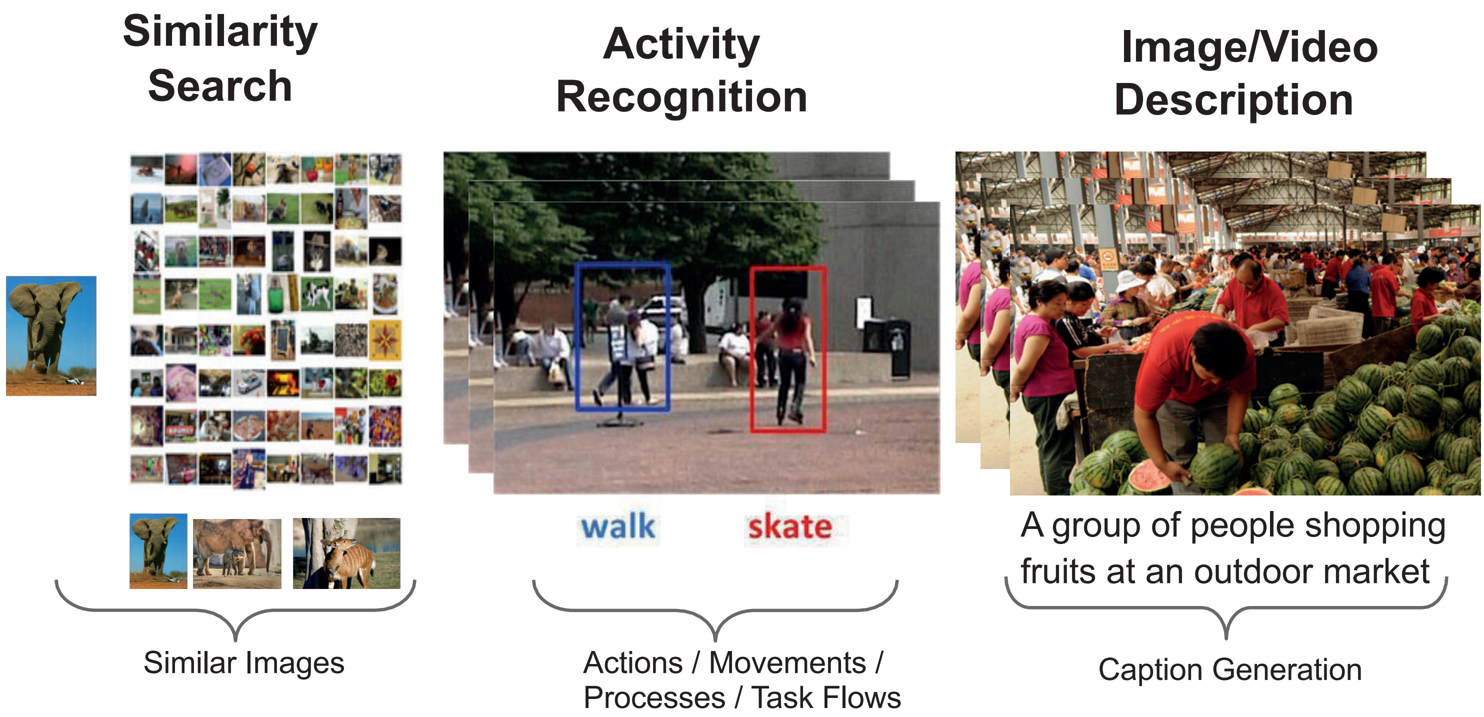
Script, Video OCR, Logos
License Plates, Ad Banners

Concept Detection

Sunset, Sea, Beach Walk

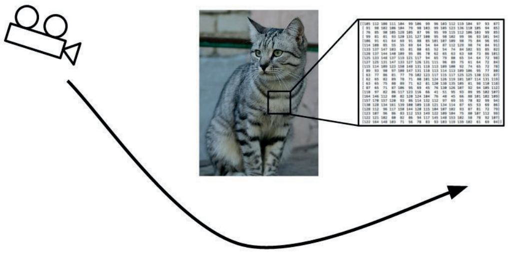
Semantic Concepts

Visual Computing Tasks



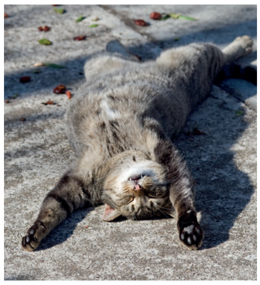
Challenges

Viewpoint



Illumination

Deformation



Occlusion

Clutter

Intraclass Variation

Challenges: Muffin or Chihuahua?

DL4VC@Marburg

Background

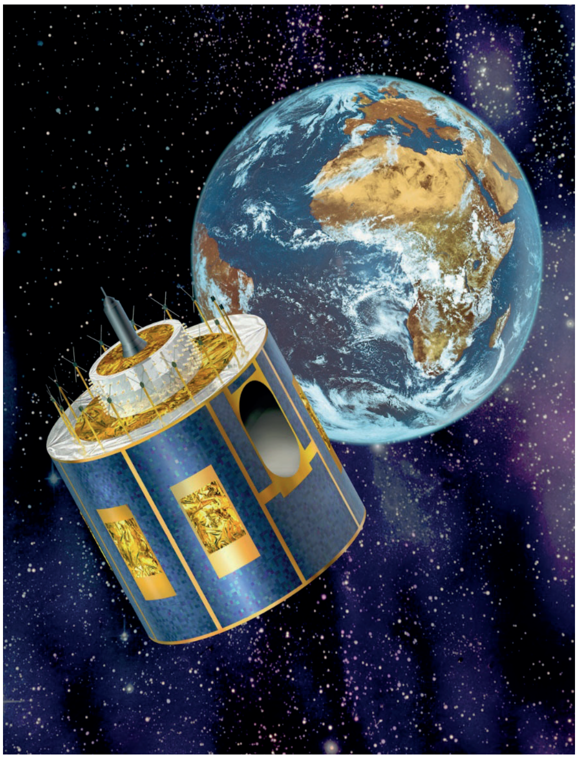
- **DFG-Project** „Content-based Image and Video Search“
  - SFB/FK 615: 2002-2010
  - PAK 509: 2010-2012
- **BMBF-Project** „MediaGrid: Distributed Analysis of Media Data“: 2009-2012
- **BMWI-Project** „Cloud-based Software Services for Semantic Search in Images and Videos“: 2011-2014
- **DFG-Project** „Content-based Search in Videos in the German Broadcast Archive“: 2012-2015 and 2018-2020
- **BMWI-Project „**GoVideo – Automatic Annotation of Documentary Film- and Video Material“: 2014-2016
- **BMBF-Project** „Florida – A Flexible System for Analyzing Video Mass Data“: 2016-2019

DL4VC@Marburg

Semantic Segmentation

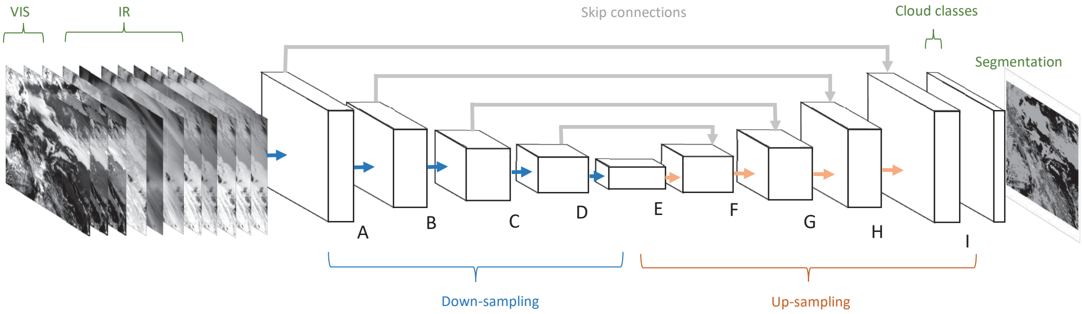
Cloud Segmentation in Satellite Images

- Cloud impact: traffic, climate, water supply...
- Meteosat Second Generation (MSG) geostationary satellite
- Spinning Enhanced Visible and Infrared Imager (SEVIRI)
  - 12 Channels
    - 3 VISual (RGB)
    - 8 InfraRed (IR)
    - 1 Panchromatic visual
  - Temporal resolution: 15 min
    → 96 scenes / day
  - Mission start: 2004
  - Spatial resolution: 3 km x 3 km (3712 x 3712 Pixel)

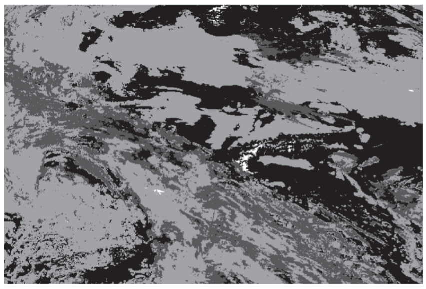


Cloud Segmentation

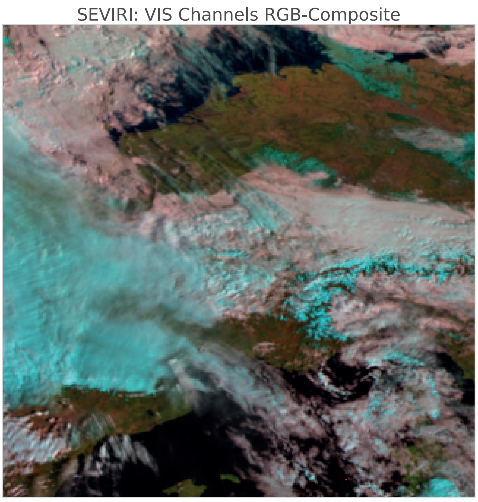
- CNN based on U-Net

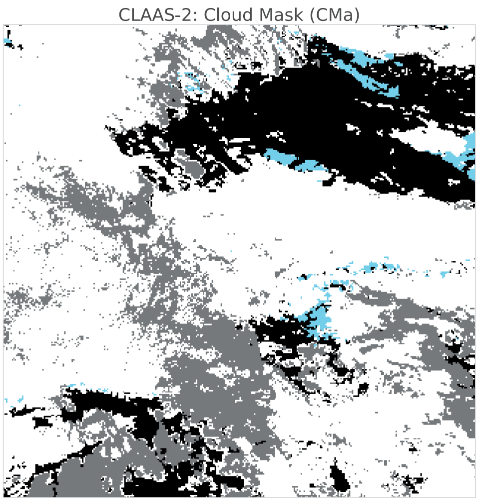


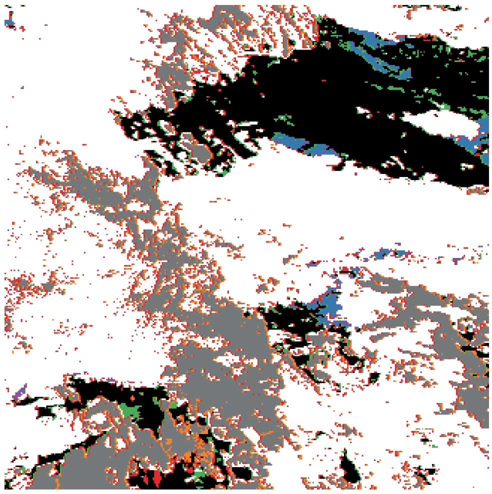
- 7, 8, or 11 channels; data (for Europe: 508 x 508 pixels)
  - Training: ~ 205000 images (2004 – 2010); test: ~ 35000 images (2012)
- Ground truth: Cloud mask from CM-SAF CMA Product
  - Manually generated decision tree from SEVIRI data – 70 pages
- Results (Accuracy)
  - Cloud free 96.0%
  - Cloud contaminated 98.6%
  - Cloud covered 94.8%
  - Snow 99.9%



Example Result







VIS

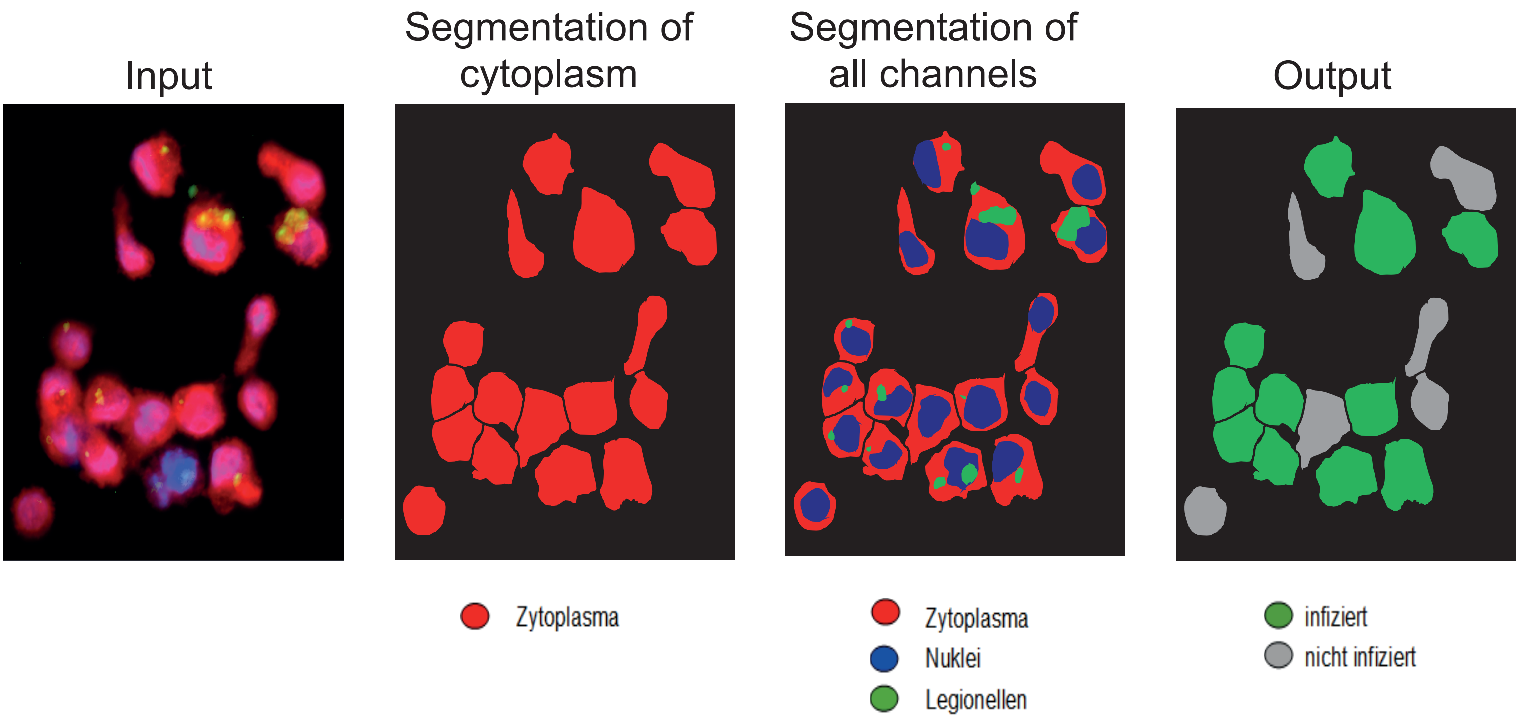
Groundtruth

Segmentation

Cloud-free = black, cloud-contaminated = gray, cloud-covered = white, snow/ice = blue

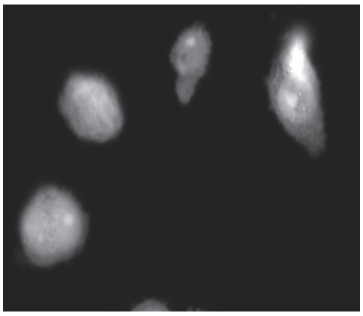
Cell Segmentation in Flourescence Microscopy Images

Aim: Determining Cells with Legionella Infestation

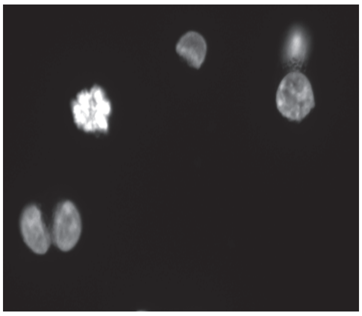


Challenges

Cytoplasm



Nuclei



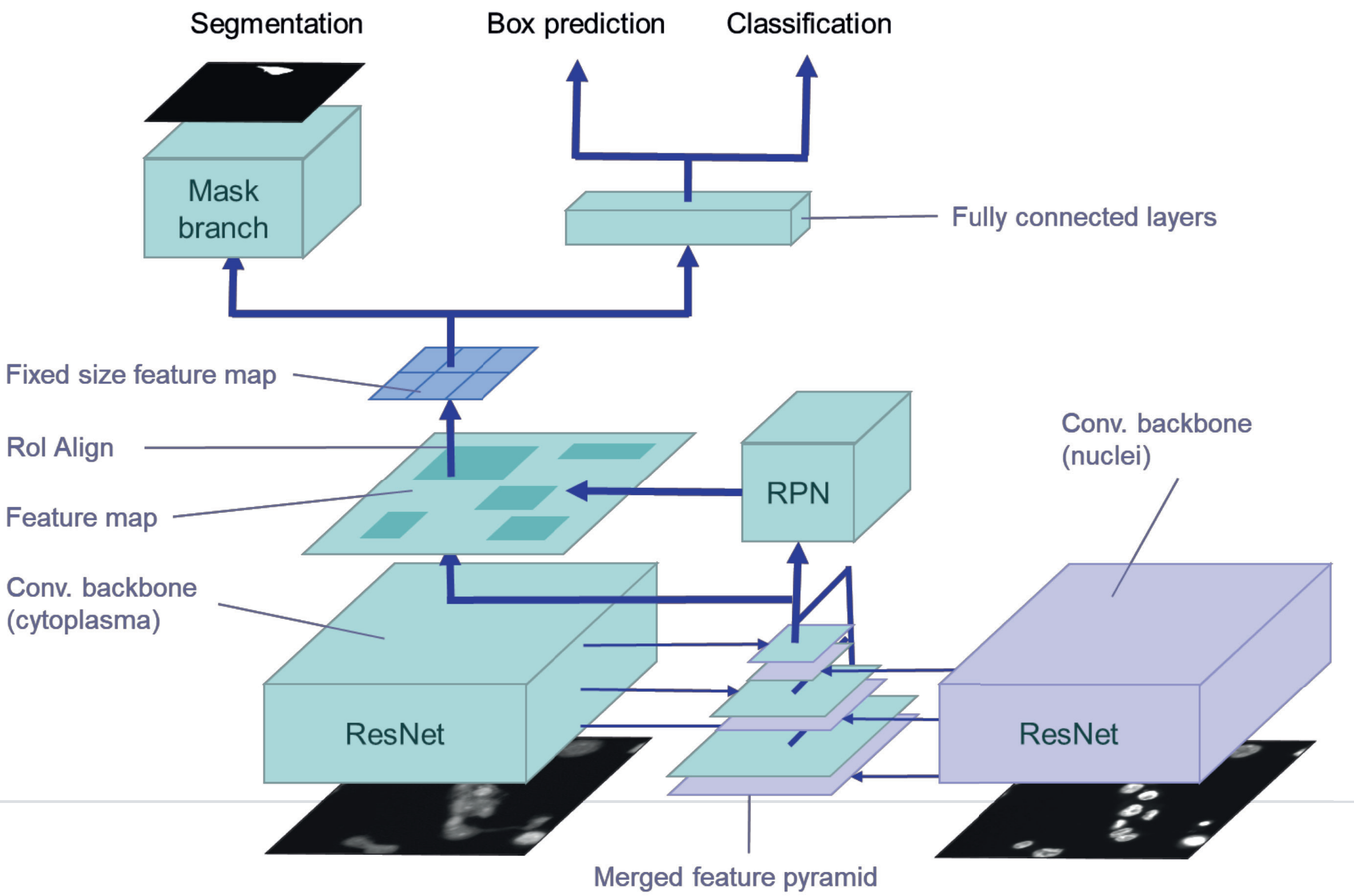
- Segmentation of the cytoplasm
- Correct separation of cells often only possible based on cell nuclei
- Only few labeled training examples (manual segmentation = high effort)

➔ Data augmentation

➔ Bounding box based segmentation (per cell/nucleus)

➔ Extended Mask-R-CNN architecture

Feature Pyramid Fusion Network

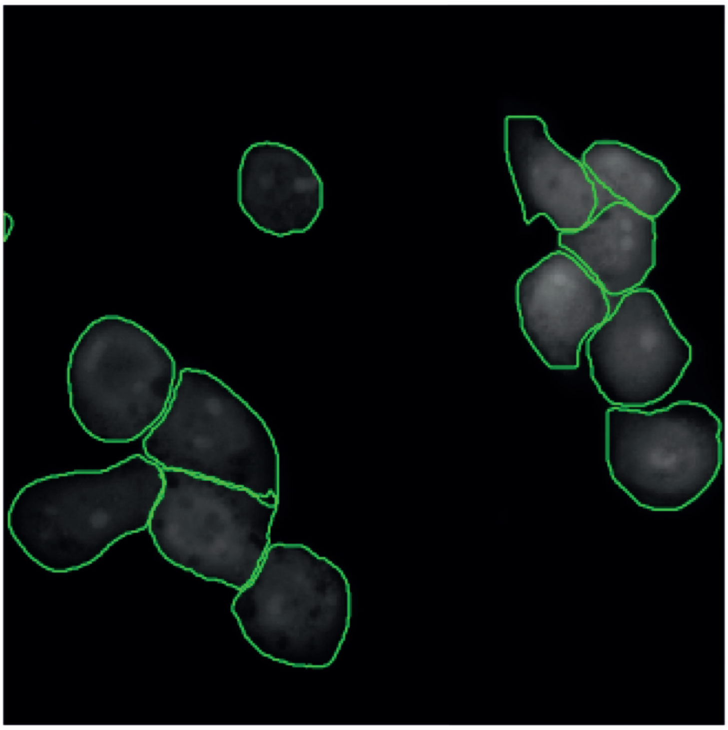


Example Result: Detection

Cells

Cells + Nuclei

Example Result: Cell Segmentation

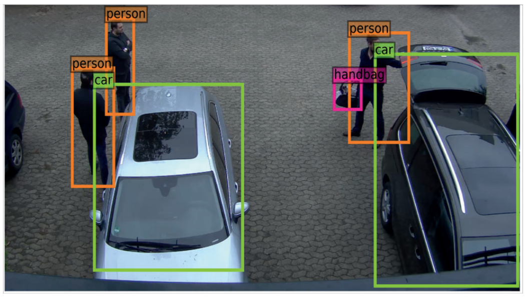


DL4VC@Marburg

Object Detection

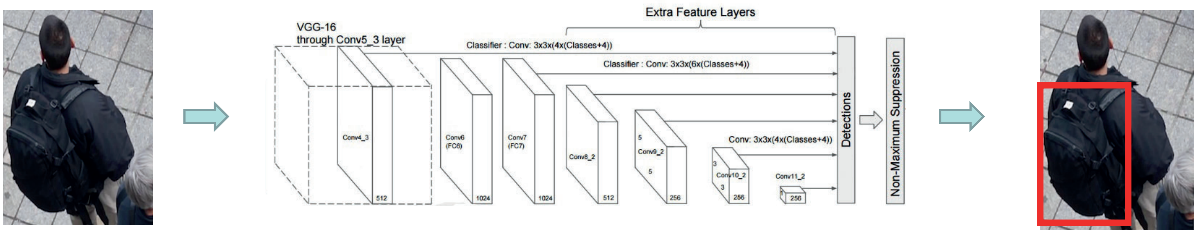
Object Detection in Surveillance Videos

- **Object classes**:
  - **Means of transport**: car, motorcycle, truck, bicycle, bus, train...
  - **Luggage**: suitcase, backpack, handbag,...
  - **Clothes:** T-shirt, jeans, coat, shirt, blazer, hat,...
  - **Animals**: dog, horse...
  - **People**: person
  - **Car license plates**
  - **UAVs (drones)**
  - **Car models**: VW Golf, 1er BMW, Renault Twingo, Audi A8...



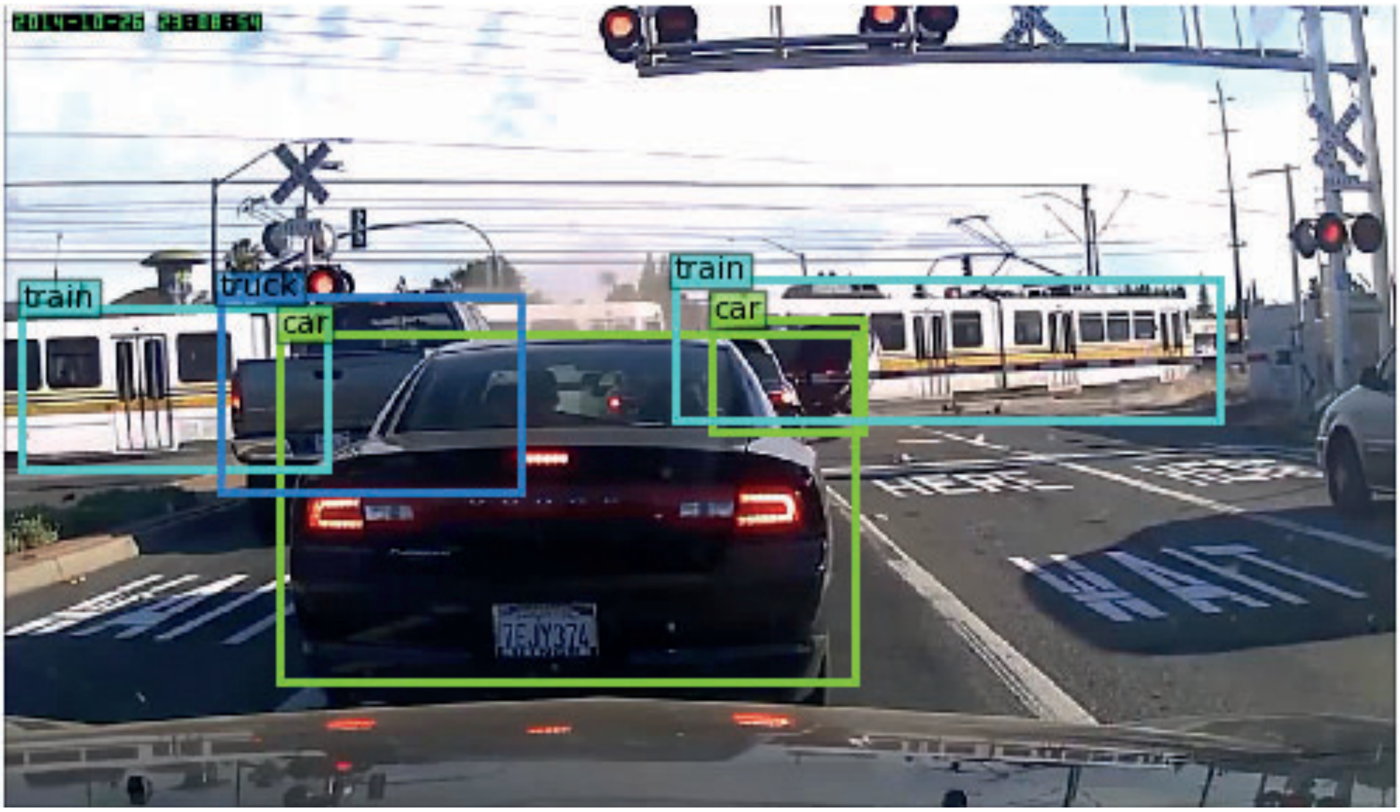
Object Detection in Surveillance Videos

- Single-Shot Multi-Box Detector (SSD) [Liu et al. 2016]



- Basic model: VGG-16 pretrained on MS COCO
- VGG-16 better than GoogLeNet and ResNet
- SSD 4x faster than Faster Region-based CNN (Faster R-CNN)
- Fine-tuning on surveillance data set (18 h)
  - Training: 2108 images with 31311 objects
- Test (challenges: small objects, motion blur, compression artifacts)
  - 56 surveillance videos (18 h):
  - 2683 images, 31506 objects

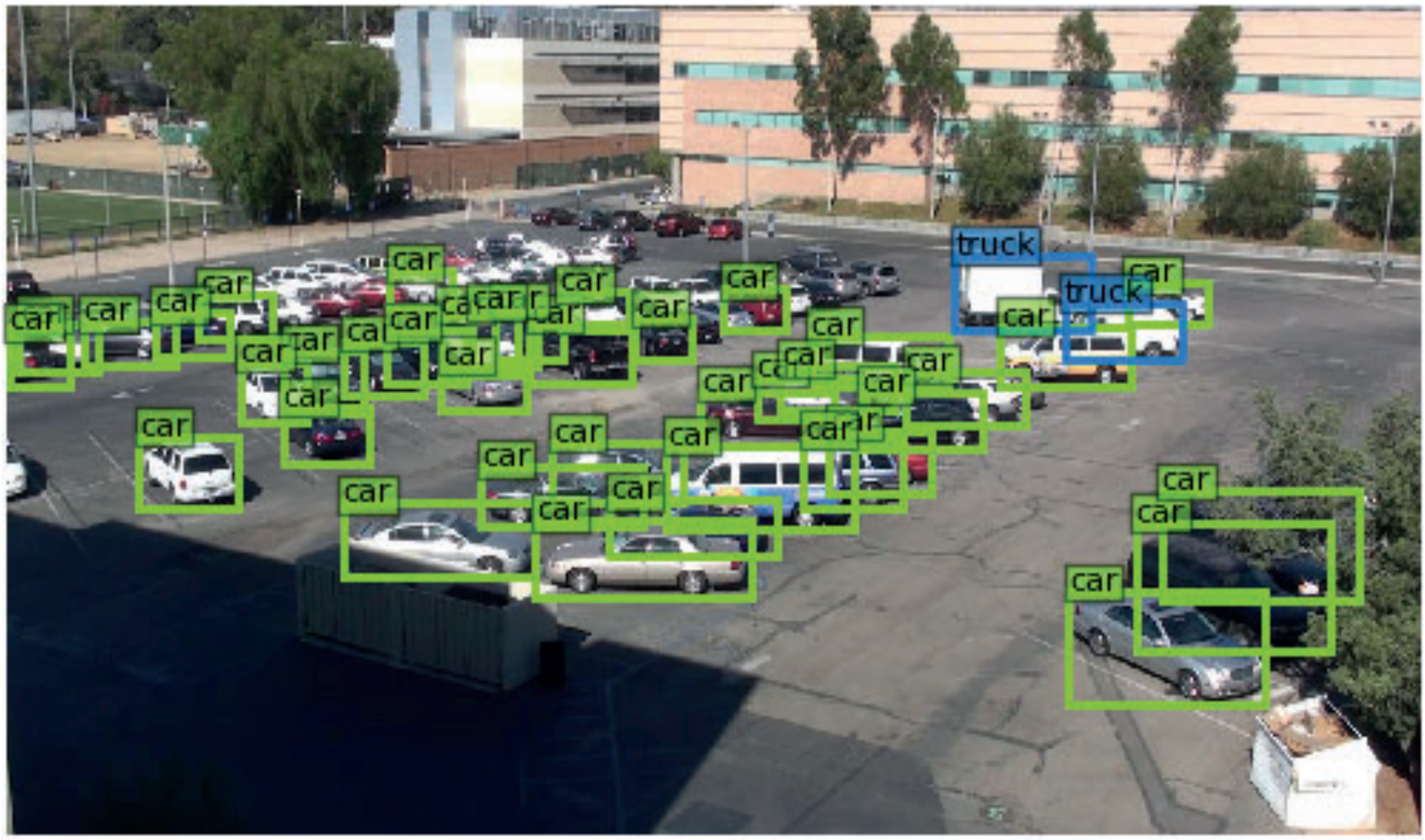
Example Results I



Example Results II

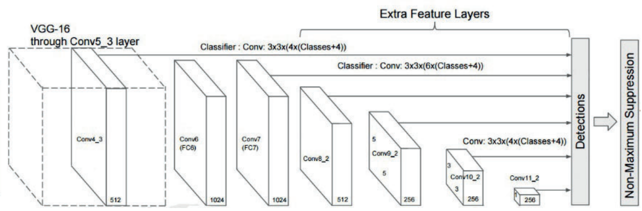
Example Results III

Example Results IV



License Plate Detection

- Single-Shot Multi-Box Detector [Liu et al. 2016]



- Basic Model: VGG-16 pretrained on IMAGENET
- Fine-tuning on license plate data set
  - Training: 4224 images with 7351 license plates
  - Validation: 377 images with 634 license plates
- Test
  - OpenALPR benchmark, MRSCORI dataset
  - 638 images with 682 license plates
  - Detection quality: 98.6% AP (Europe), 98.3% AP (USA)

Example: License Plate Detection

Example: Car Model Recognition

- Data acquisition: Webcrawler
- Spam filtering

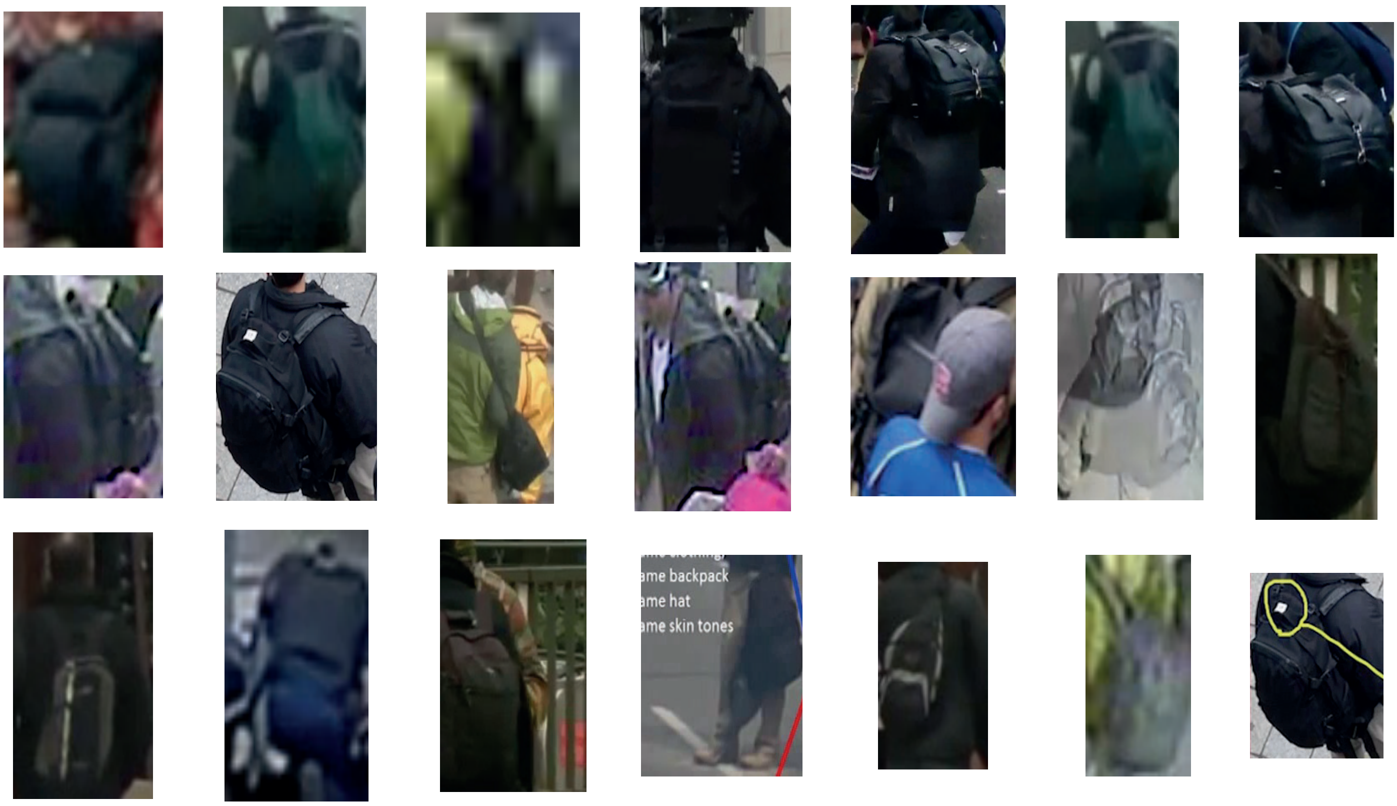
Spam

Ham

- Data
  - 2,202,842 training images
  - 74 car makes, 835 car models
  - 50 test images per model
- Network architecture: Mobile NASNet
- Example video: https://box.uni-marburg.de/index.php/s/UCNcGanjysU2qHD

Example: Youtube Videos – Knapsack Retrieval

8519 Videos (Berlin, Boston, Dallas, Istanbul, London, Nizza, Paris)



DL4VC@Marburg

Concept Detection

Person Recognition

Text Spotting / Video OCR

German Broadcasting Archive (DRA)

- Founded in 1952
- Charitable foundation and joint institution of the ARD
- Historical collections of scientifically relevant videos
- Cultural heritage of GDR TV broadcasts
  - ~ 100,000 broadcasts (1952 – 1991)
    - Daily news program „Aktuelle Kamera“
    - Political magazines (e.g., „Prisma“)
    - Films, film adaptations and TV series (e.g. „Polizeiruf 110”)
    - Entertainment programs (e.g., „Ein Kessel Buntes“)
    - Children‘s and youth programs
    - Advice and sports programs
  - Considerable research interest in GDR and German-German history

Concept Lexicon

- Based on analysis of user search queries
- Focus on queries that are difficult and time-consuming to answer
- 100 GDR-specific concepts
  - Scenes or places
    - Optical industry, supermarket, railroad station, daylight mine, production hall, camping site, kindergarten, shopping hall, kitchen, allotment, …
  - Events or activities
    - Border control, concert, applauding, handshake, brotherly kiss, wreath ceremony..
  - Objects
    - Trabant, GDR emblem, ambulance, GDR flag, tram, German state railway, …
  - Persons
    - Teenager, “Abschnittsbevollmächtiger”, …
  - Personalities
    - Erich Honecker, Walter Ulbricht, Hilde Benjamin, Siegmund Jähn, …

Dataset

- Historical GDR television recordings
- Technically very challenging
  - Many recordings are grayscale
  - Low technical quality (the older, the poorer the video quality)

Training data

- 416,249 video shots
- 118,020 annotated video frames
- 91 concepts (77 evaluated)
- 9 persons

Test data

- 1,545,600 video shots
- ~ 2490 h videos

Concept Detection Examples

Militär-
parade

Schlot

Plattenbau

Straßen-
verkehr

Person Recognition Results

Erich Honecker

Christa Wolf

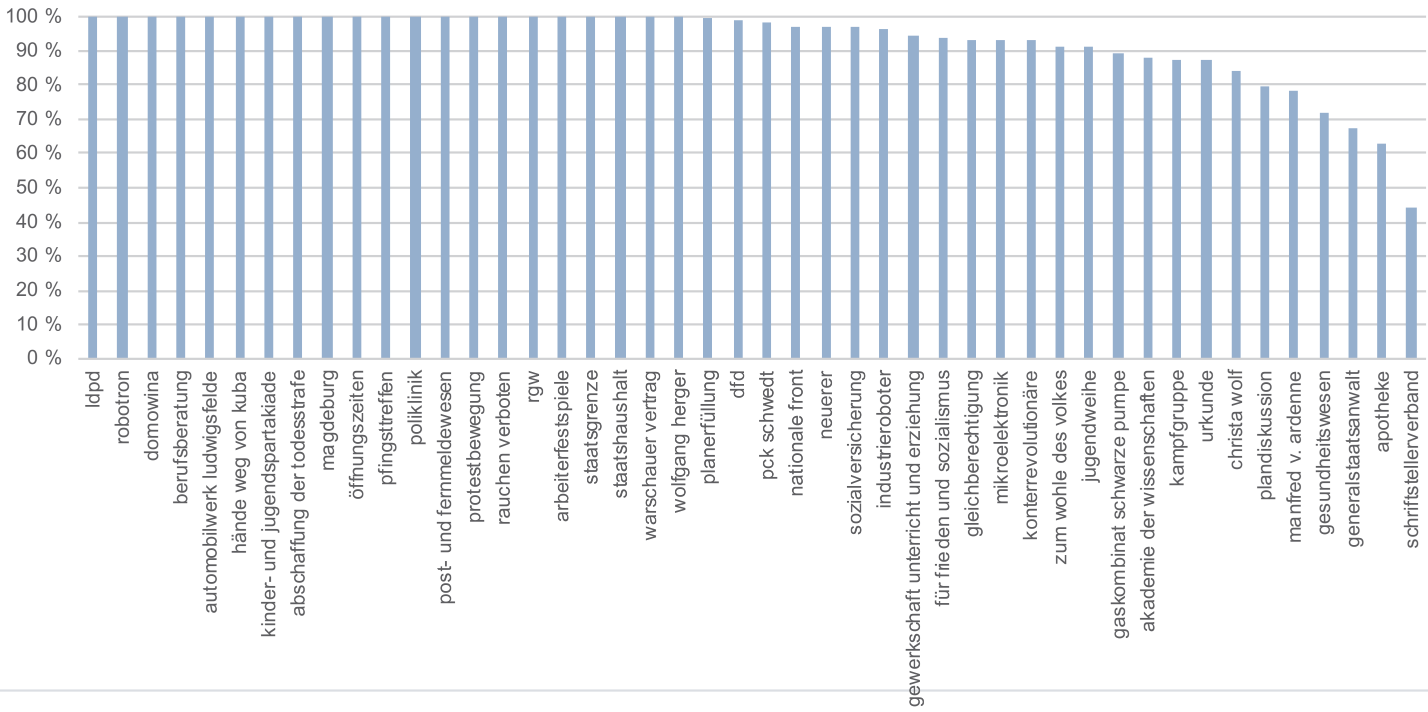
Walter Ulbricht

Hilde Benjamin

Video OCR Results

- 46 text queries, evaluation based on the top-100 results per query

  => 92.9% Mean Average Precision



DL4VC@Marburg

Similarity Search

What is Similarity?

- Semantic vs. pixel based similarity

- Fine-grained image similarity

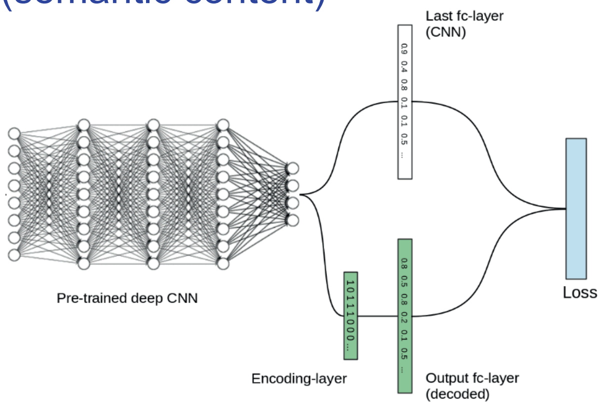
  

- Similar?

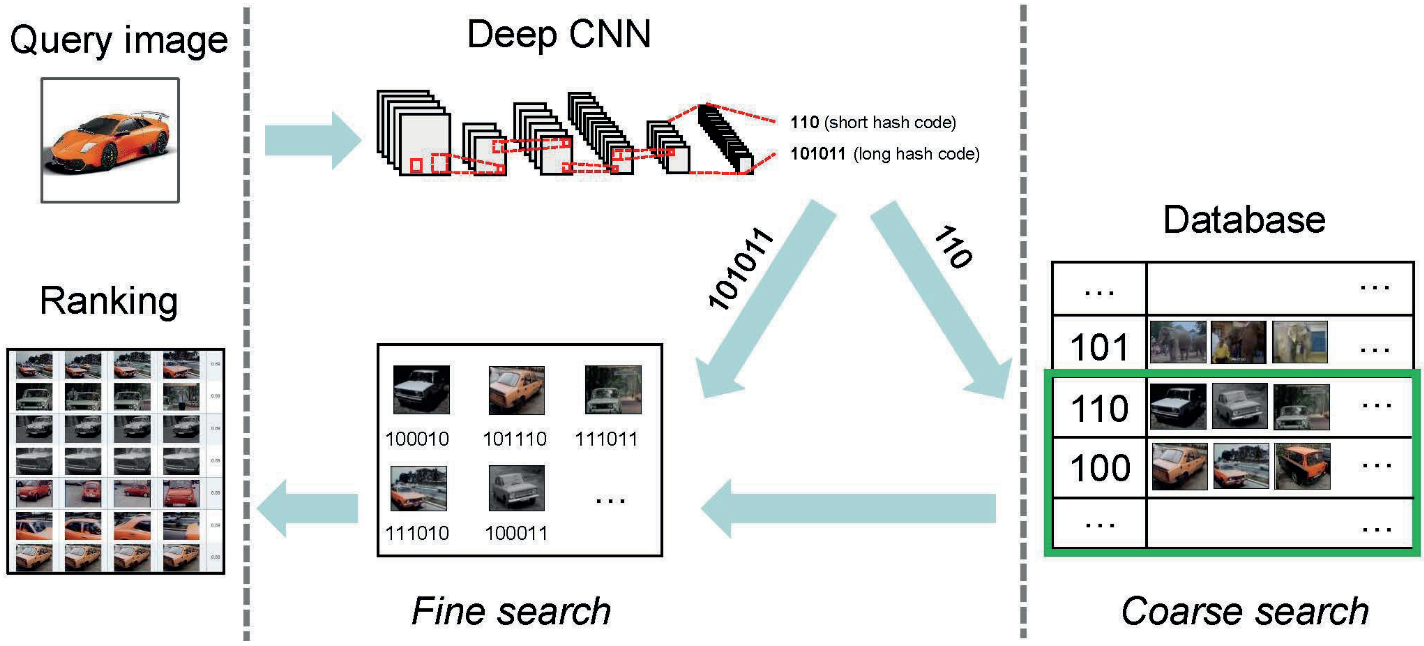
Similarity Search

- Query by example
- Features based on CNNs
  - Better suited for objects and scenes (semantic content)
  - Less dependent on pixel intensities
- Semantic hashing
  - Learning binary codes for images
  - Compact representation
  - Fast matching

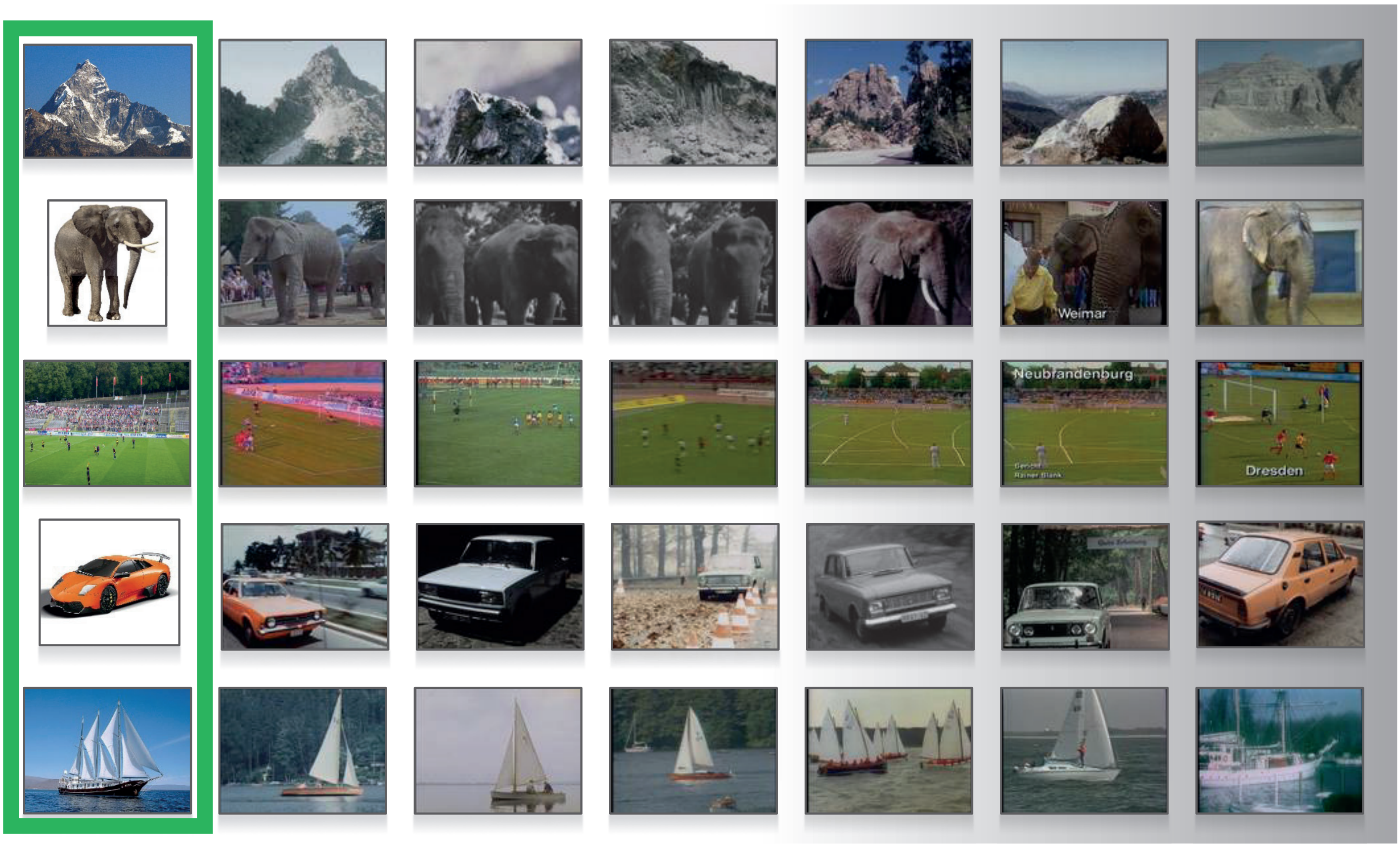


- Two stage approach
  - Coarse-level search based on 64 bit binary codes using a Vantage-Point tree
    - ➔ „Short" list of potential results
  - Fine-level search with 256 bit codes based on the short list

Similarity Search: Semantic Hashing



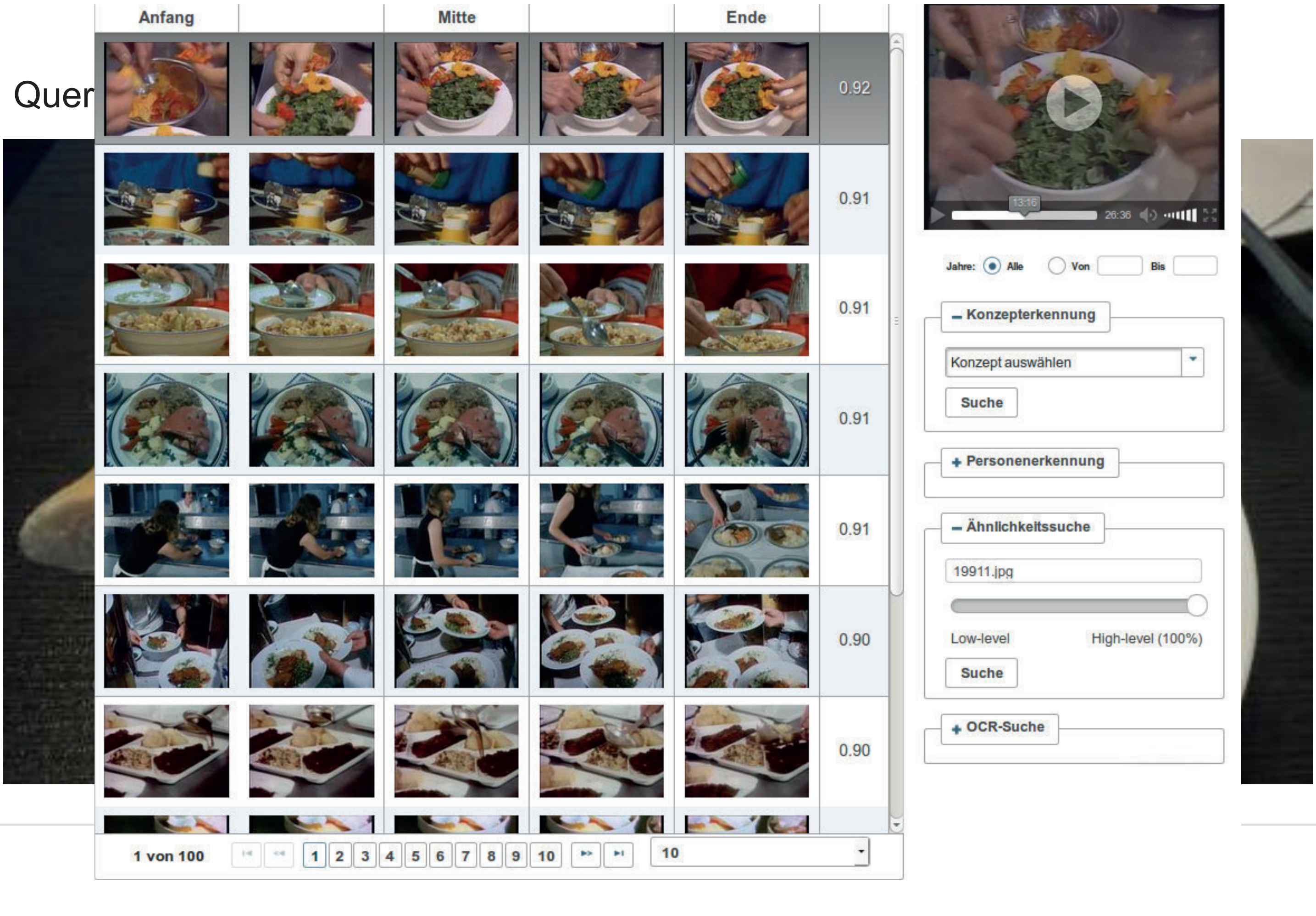
Similarity Search Results



1st column: query images downloaded from the WWW

Similarity Search Results

Quer



Conclusion

- Deep learning = Learning Hierarchical Representations
- Deep learning is highly promising for *visual computing (*but also for *audio processing, sensor processing,* and *natural language processing)*
- Current & future work:
  - Anomaly detection in surveillance cameras of chemical process plants
  - Deep learning for e-health / m-health applications
  - Deep learning on mobile devices (Qualcomm 835, Nvidia Jetson TX2)
  - Unsupervised deep learning for network traffic analysis ("packet analytics")
  - Deep reinforcement learning for robotics (UAVs, UGVs, coordination…)
  - Deep learning for sequential data / streams (music, text, clickstreams…)

Slide / Figure Credits

- Markus Mühling, University of Marburg, Germany
- Yousri Kessentini, University of Sfax, Tunisia
- Fei-Fei Lee, Stanford University, USA
- Bart ter Haar Romeny, Eindhoven University of Technology, The Netherlands
- Weifeng Lee et al., University of Arizona, USA
- Qiang Yang, Hongkong University of Science and Technology, China