

Computer Vision - Lecture 15

Deep Learning for Object Categorization

21.12.2016

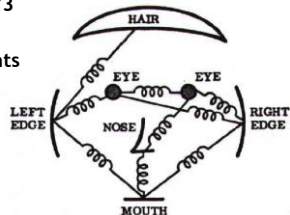
Bastian Leibe
 RWTH Aachen
<http://www.vision.rwth-aachen.de>
 leibe@vision.rwth-aachen.de

Course Outline

- Image Processing Basics
- Segmentation & Grouping
- Object Recognition
- Object Categorization I
 - Sliding Window based Object Detection
- Local Features & Matching
 - Local Features - Detection and Description
 - Recognition with Local Features
 - Indexing & Visual Vocabularies
- Object Categorization II
 - Bag-of-Words Approaches & Part-based Approaches
 - Deep Learning Methods
- 3D Reconstruction

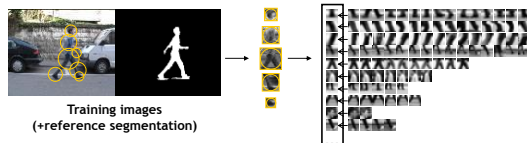
Recap: Part-Based Models

- Fischler & Elschlager 1973
- Model has two components
 - parts (2D image fragments)
 - structure (configuration of parts)

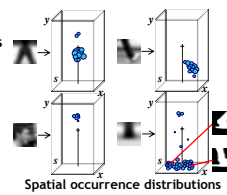


B. Leibe

Recap: Implicit Shape Model - Representation

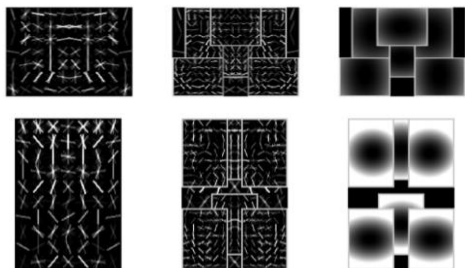


- Learn appearance codebook
 - Extract local features at interest points
 - Clustering \Rightarrow appearance codebook
- Learn spatial distributions
 - Match codebook to training images
 - Record matching positions on object



B. Leibe

Recap: Deformable Part-Based Model



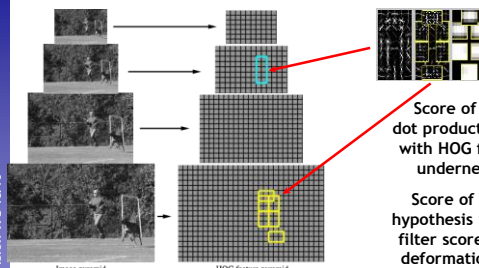
Root filters
coarse resolution

Part filters
finer resolution

Deformation models

B. Leibe

Recap: Object Hypothesis



Score of filter:
dot product of filter with HOG features underneath it

Score of object hypothesis is sum of filter scores minus deformation costs

- Multiscale model captures features at two resolutions

B. Leibe

RWTH AACHEN UNIVERSITY

Recap: Score of a Hypothesis

$$\text{score}(p_0, \dots, p_n) = \sum_{i=0}^n F_i \cdot \phi(H, p_i) - \sum_{i=1}^n d_i \cdot (dx_i^2, dy_i^2)$$

"data term"

↑

filters

"spatial prior"

↑

displacements

deformation parameters

$\text{score}(z) = \beta \cdot \Psi(H, z)$

concatenation filters and deformation parameters

concatenation of HOG features and part displacement features

Computer Vision WS 15/16 8

Slide credit: Pedro Felzenszwalb B. Leibe

RWTH AACHEN UNIVERSITY

Topics of This Lecture

- **Deep Learning**
 - Motivation
- **Convolutional Neural Networks**
 - Convolutional Layers
 - Pooling Layers
 - Nonlinearities
- **CNN Architectures**
 - LeNet
 - AlexNet
 - VGGNet
 - GoogLeNet
- **Applications**

Computer Vision WS 15/16 9

B. Leibe

RWTH AACHEN UNIVERSITY

We've finally got there!

Deep Learning

Computer Vision WS 15/16 10

B. Leibe

RWTH AACHEN UNIVERSITY

Traditional Recognition Approach

Image/
Video
Pixels

Hand-designed
feature
extraction

➔

Trainable
classifier

➔

Object
Class

- **Characteristics**
 - Features are not learned, but engineered
 - Trainable classifier is often generic (e.g., SVM)

⇒ Many successes in 2000-2010.

Computer Vision WS 15/16 11

Slide credit: Svetlana Lazebnik B. Leibe

RWTH AACHEN UNIVERSITY

Traditional Recognition Approach

- Features are key to recent progress in recognition
 - Multitude of hand-designed features currently in use
 - SIFT, HOG,

⇒ *Where next? Better classifiers? Or keep building more features?*

DPM
 [Felzenszwalb et al., PAMI'07]

Dense SIFT+LBP+HOG → BOW → Classifier
 [Yan & Huan '10]
 (Winner of PASCAL 2010 Challenge)

Computer Vision WS 15/16 12

Slide credit: Svetlana Lazebnik B. Leibe

RWTH AACHEN UNIVERSITY

What About Learning the Features?

- Learn a *feature hierarchy* all the way from pixels to classifier
 - Each layer extracts features from the output of previous layer
 - Train all layers jointly

Image/
Video
Pixels

Layer 1

➔

Layer 2

➔

Layer 3

➔

Simple
Classifier

Computer Vision WS 15/16 13

Slide credit: Svetlana Lazebnik B. Leibe

RWTH AACHEN UNIVERSITY

“Shallow” vs. “Deep” Architectures

Traditional recognition: “Shallow” architecture

Deep learning: “Deep” architecture

Computer Vision WS 15/16 14

Slide credit: Svetlana Lazebnik B. Leibe

RWTH AACHEN UNIVERSITY

Background: Perceptrons

Input

Weights

x_1
 x_2
 x_3
 \vdots
 x_d

w_1
 w_2
 w_3
 \vdots
 w_d

Output: $\sigma(\mathbf{w} \cdot \mathbf{x} + b)$

Sigmoid function

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

Computer Vision WS 15/16 15

Slide credit: Svetlana Lazebnik

RWTH AACHEN UNIVERSITY

Inspiration: Neuron Cells

Computer Vision WS 15/16 16

Slide credit: Svetlana Lazebnik, Rob Fergus

RWTH AACHEN UNIVERSITY

Background: Multi-Layer Neural Networks

- **Nonlinear classifier**
 - **Training:** find network weights \mathbf{w} to minimize the error between true training labels t_n and estimated labels $f_{\mathbf{w}}(x_n)$:

$$E(\mathbf{W}) = \sum_n L(t_n, f(\mathbf{x}_n; \mathbf{W}))$$
 - Minimization can be done by gradient descent, provided f is differentiable
 - Training method: **error backpropagation.**

Computer Vision WS 15/16 17

Slide credit: Svetlana Lazebnik B. Leibe

RWTH AACHEN UNIVERSITY

Hubel/Wiesel Architecture

- **D. Hubel, T. Wiesel (1959, 1962, Nobel Prize 1981)**
 - Visual cortex consists of a hierarchy of *simple*, *complex*, and *hyper-complex* cells

Hubel & Wiesel

topographical mapping

featural hierarchy

high level

mid level

low level

Computer Vision WS 15/16 18

Slide credit: Svetlana Lazebnik, Rob Fergus B. Leibe

RWTH AACHEN UNIVERSITY

Convolutional Neural Networks (CNN, ConvNet)

- **Neural network with specialized connectivity structure**
 - Stack multiple stages of feature extractors
 - Higher stages compute more global, more invariant features
 - Classification layer at the end

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, [Gradient-based learning applied to document recognition](#), Proceedings of the IEEE 86(11): 2278-2324, 1998.

Computer Vision WS 15/16 19

Slide credit: Svetlana Lazebnik B. Leibe

RWTH AACHEN UNIVERSITY

Topics of This Lecture

- Deep Learning
 - Motivation
- Convolutional Neural Networks
 - Convolutional Layers
 - Pooling Layers
 - Nonlinearities
- CNN Architectures
 - LeNet
 - AlexNet
 - VGGNet
 - GoogLeNet
- Applications

Computer Vision WS 15/16 20

RWTH AACHEN UNIVERSITY

Convolutional Networks: Structure

- Feed-forward feature extraction
 1. Convolve input with learned filters
 2. Non-linearity
 3. Spatial pooling
 4. (Normalization)
- Supervised training of convolutional filters by back-propagating classification error

Computer Vision WS 15/16 21

RWTH AACHEN UNIVERSITY

Convolutional Networks: Intuition

- Fully connected network
 - E.g. 1000×1000 image
 - 1M hidden units
 - ⇒ 1T parameters!
- Ideas to improve this
 - Spatial correlation is local

Computer Vision WS 15/16 22

RWTH AACHEN UNIVERSITY

Convolutional Networks: Intuition

- Locally connected net
 - E.g. 1000×1000 image
 - 1M hidden units
 - 10×10 receptive fields
 - ⇒ 100M parameters!
- Ideas to improve this
 - Spatial correlation is local
 - Want translation invariance

Computer Vision WS 15/16 23

RWTH AACHEN UNIVERSITY

Convolutional Networks: Intuition

- Convolutional net
 - Share the same parameters across different locations
 - Convolutions with learned kernels

Computer Vision WS 15/16 24

RWTH AACHEN UNIVERSITY

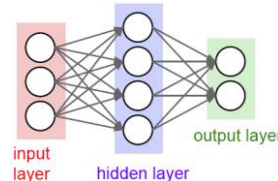
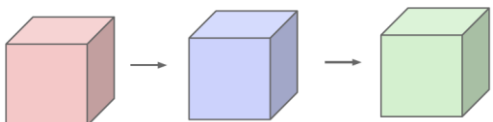
Convolutional Networks: Intuition

- Convolutional net
 - Share the same parameters across different locations
 - Convolutions with learned kernels
- Learn multiple filters
 - E.g. 1000×1000 image
 - 100 filters
 - 10×10 filter size
 - ⇒ 10k parameters
- Result: Response map
 - size: 1000×1000×100
 - Only memory, not params!

Computer Vision WS 15/16 25

RWTH AACHEN UNIVERSITY

Important Conceptual Shift

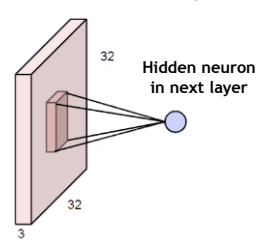
- Before
 
- Now:
 

Computer Vision WS 15/16 26

Slide credit: FeiFei Li, Andrei Karpathy B. Leibe

RWTH AACHEN UNIVERSITY

Convolution Layers



Example image: $32 \times 32 \times 3$ volume

Before: Full connectivity $32 \times 32 \times 3$ weights

Now: Local connectivity
One neuron connects to, e.g., $5 \times 5 \times 3$ region.
⇒ Only $5 \times 5 \times 3$ shared weights.

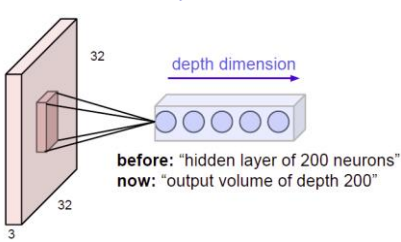
- Note: Connectivity is
 - Local in space (5×5 inside 32×32)
 - But full in depth (all 3 depth channels)

Computer Vision WS 15/16 27

Slide adapted from FeiFei Li, Andrei Karpathy B. Leibe

RWTH AACHEN UNIVERSITY

Convolution Layers



before: "hidden layer of 200 neurons"
now: "output volume of depth 200"

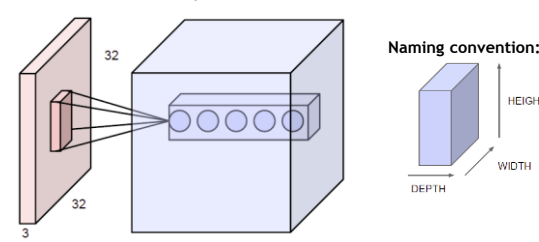
- All Neural Net activations arranged in 3 dimensions
 - Multiple neurons all looking at the same input region, stacked in depth

Computer Vision WS 15/16 28

Slide adapted from FeiFei Li, Andrei Karpathy B. Leibe

RWTH AACHEN UNIVERSITY

Convolution Layers



Naming convention:

HEIGHT
WIDTH
DEPTH

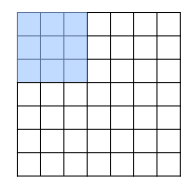
- All Neural Net activations arranged in 3 dimensions
 - Multiple neurons all looking at the same input region, stacked in depth
 - Form a single $[1 \times 1 \times \text{depth}]$ depth column in output volume.

Computer Vision WS 15/16 29

Slide credit: FeiFei Li, Andrei Karpathy B. Leibe

RWTH AACHEN UNIVERSITY

Convolution Layers



Example:
 7×7 input
assume 3×3 connectivity
stride 1

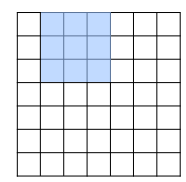
- Replicate this column of hidden neurons across space, with some stride.

Computer Vision WS 15/16 30

Slide credit: FeiFei Li, Andrei Karpathy B. Leibe

RWTH AACHEN UNIVERSITY

Convolution Layers



Example:
 7×7 input
assume 3×3 connectivity
stride 1

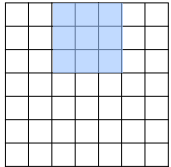
- Replicate this column of hidden neurons across space, with some stride.

Computer Vision WS 15/16 31

Slide credit: FeiFei Li, Andrei Karpathy B. Leibe

RWTH AACHEN UNIVERSITY

Convolution Layers



Example:
7x7 input
assume 3x3 connectivity
stride 1

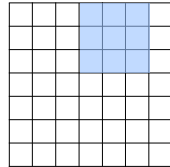
- Replicate this column of hidden neurons across space, with some **stride**.

Computer Vision WS 15/16 32

Slide credit: FeiFei Li, Andrei Karpathy B. Leibe

RWTH AACHEN UNIVERSITY

Convolution Layers



Example:
7x7 input
assume 3x3 connectivity
stride 1

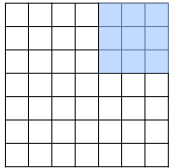
- Replicate this column of hidden neurons across space, with some **stride**.

Computer Vision WS 15/16 33

Slide credit: FeiFei Li, Andrei Karpathy B. Leibe

RWTH AACHEN UNIVERSITY

Convolution Layers



Example:
7x7 input
assume 3x3 connectivity
stride 1
⇒ 5x5 output

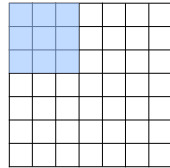
- Replicate this column of hidden neurons across space, with some **stride**.

Computer Vision WS 15/16 34

Slide credit: FeiFei Li, Andrei Karpathy B. Leibe

RWTH AACHEN UNIVERSITY

Convolution Layers



Example:
7x7 input
assume 3x3 connectivity
stride 1
⇒ 5x5 output

What about stride 2?

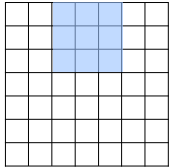
- Replicate this column of hidden neurons across space, with some **stride**.

Computer Vision WS 15/16 35

Slide credit: FeiFei Li, Andrei Karpathy B. Leibe

RWTH AACHEN UNIVERSITY

Convolution Layers



Example:
7x7 input
assume 3x3 connectivity
stride 1
⇒ 5x5 output

What about stride 2?

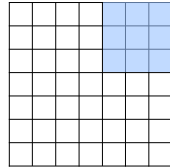
- Replicate this column of hidden neurons across space, with some **stride**.

Computer Vision WS 15/16 36

Slide credit: FeiFei Li, Andrei Karpathy B. Leibe

RWTH AACHEN UNIVERSITY

Convolution Layers



Example:
7x7 input
assume 3x3 connectivity
stride 1
⇒ 5x5 output

What about stride 2?
⇒ 3x3 output

- Replicate this column of hidden neurons across space, with some **stride**.

Computer Vision WS 15/16 37

Slide credit: FeiFei Li, Andrei Karpathy B. Leibe

RWTH AACHEN UNIVERSITY

Convolution Layers

0	0	0	0	0		
0						
0						
0						
0						

Example:
 7x7 input
 assume 3x3 connectivity
 stride 1
 ⇒ 5x5 output

What about stride 2?
 ⇒ 3x3 output

- Replicate this column of hidden neurons across space, with some **stride**.
- In practice, common to zero-pad the border.
 - Preserves the size of the input spatially.

Computer Vision WS 15/16 38

Slide credit: FeiFei Li, Andrei Karpathy. B. Leibe

RWTH AACHEN UNIVERSITY

Activation Maps of Convolutional Filters

Activations: one filter = one depth slice (or activation map)

5x5 filters

Each activation map is a depth slice through the output volume.

Activation maps

Computer Vision WS 15/16 39

Slide adapted from FeiFei Li, Andrei Karpathy. B. Leibe

RWTH AACHEN UNIVERSITY

Effect of Multiple Convolution Layers

Low-Level Feature → Mid-Level Feature → High-Level Feature → Trainable Classifier

Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

Computer Vision WS 15/16 40

Slide credit: Yann LeCun. B. Leibe

RWTH AACHEN UNIVERSITY

Commonly Used Nonlinearities

- Sigmoid**

$$g(a) = \sigma(a) = \frac{1}{1 + \exp\{-a\}}$$
- Hyperbolic tangent**

$$g(a) = \tanh(a) = 2\sigma(2a) - 1$$
- Rectified linear unit (ReLU)**

$$g(a) = \max\{0, a\}$$

Preferred option for deep networks

Computer Vision WS 15/16 41

B. Leibe

RWTH AACHEN UNIVERSITY

Convolutional Networks: Intuition

- Let's assume the filter is an eye detector
 - How can we make the detection robust to the exact location of the eye?

Computer Vision WS 15/16 42

Slide adapted from Marc'Aurelio Ranzato. B. Leibe. Image source: Yann LeCun

RWTH AACHEN UNIVERSITY

Convolutional Networks: Intuition

- Let's assume the filter is an eye detector
 - How can we make the detection robust to the exact location of the eye?
- Solution:**
 - By **pooling** (e.g., max or avg) filter responses at different spatial locations, we gain robustness to the exact spatial location of features.

Computer Vision WS 15/16 43

Slide adapted from Marc'Aurelio Ranzato. B. Leibe. Image source: Yann LeCun

Computer Vision WS 15/16

Max Pooling

Single depth slice

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

max pool with 2x2 filters and stride 2

6	8
3	4

- Effect:
 - Make the representation smaller without losing too much information
 - Achieve robustness to translations

Slide adapted from FeiFei Li, Andrei Karpathy. B. Leibe

Computer Vision WS 15/16

Max Pooling

Single depth slice

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

max pool with 2x2 filters and stride 2

6	8
3	4

- Note
 - Pooling happens independently across each slice, preserving the number of slices.

Slide adapted from FeiFei Li, Andrei Karpathy. B. Leibe

Computer Vision WS 15/16

Compare: SIFT Descriptor

Image Pixels

Apply oriented filters

Spatial pool (Sum)

Normalize to unit length

Feature Vector

Lowe [IJCV 2004]

Slide credit: Svetlana Lazebnik. B. Leibe

Computer Vision WS 15/16

Compare: Spatial Pyramid Matching

SIFT features

Filter with Visual Words

Take max VW response (L-inf normalization)

Multi-scale spatial pool (Sum)

Global image descriptor

Lazebnik, Schmid, Ponce [CVPR 2006]

Slide credit: Svetlana Lazebnik. B. Leibe

Computer Vision WS 15/16

Topics of This Lecture

- Deep Learning
 - Motivation
- Convolutional Neural Networks
 - Convolutional Layers
 - Pooling Layers
 - Nonlinearities
- CNN Architectures
 - LeNet
 - AlexNet
 - VGGNet
 - GoogLeNet
- Applications

B. Leibe

Computer Vision WS 15/16

CNN Architectures: LeNet (1998)

- Early convolutional architecture
 - 2 Convolutional layers, 2 pooling layers
 - Fully-connected NN layers for classification
 - Successfully used for handwritten digit recognition (MNIST)

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, [Gradient-based learning applied to document recognition](#), Proceedings of the IEEE 86(11): 2278-2324, 1998.


Slide credit: Svetlana Lazebnik. B. Leibe

Computer Vision WS 15/16

ImageNet Challenge 2012

RWTH AACHEN UNIVERSITY

- ImageNet
 - 14M labeled internet images
 - 20k classes
 - Human labels via Amazon Mechanical Turk
- Challenge (ILSVRC)
 - 1.2 million training images
 - 1000 classes
 - Goal: Predict ground-truth class within top-5 responses
 - Currently one of the top benchmarks in Computer Vision



[Deng et al., CVPR'09]

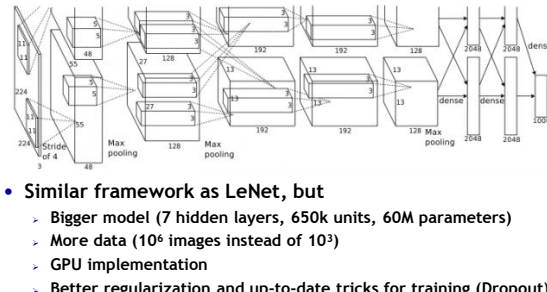
B. Leibe

50

Computer Vision WS 15/16

CNN Architectures: AlexNet (2012)

RWTH AACHEN UNIVERSITY



- Similar framework as LeNet, but
 - Bigger model (7 hidden layers, 650k units, 60M parameters)
 - More data (10^6 images instead of 10^3)
 - GPU implementation
 - Better regularization and up-to-date tricks for training (Dropout)

A. Krizhevsky, I. Sutskever, and G. Hinton, [ImageNet Classification with Deep Convolutional Neural Networks](#), NIPS 2012.

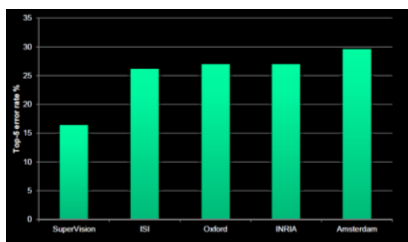
B. Leibe

51

Computer Vision WS 15/16

ILSVRC 2012 Results

RWTH AACHEN UNIVERSITY



Team	Top-5 error rate %
SuperVision	~16.4
ISI	~26.2
Oxford	~26.2
NIPS	~26.2
Amsterdam	~26.2

- AlexNet almost halved the error rate
 - 16.4% error (top-5) vs. 26.2% for the next best approach
 - ⇒ A revolution in Computer Vision
 - Acquired by Google in Jan '13, deployed in Google+ in May '13

B. Leibe

52

Computer Vision WS 15/16

AlexNet Results

RWTH AACHEN UNIVERSITY

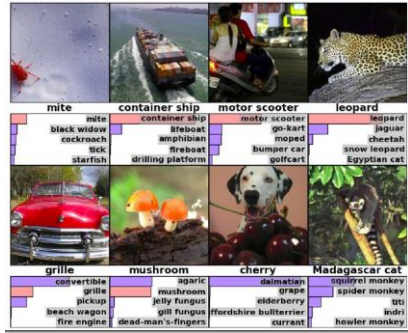


Image	Ground Truth	AlexNet Top-5 Predictions
	mite	mite, Black widow, cockroach, tick, starfish
	container ship	lifeboat, amphibian, fireboat, drilling platform
	motor scooter	go-kart, moped, bumper car, golfcart
	leopard	leopard, jaguar, cheetah, snow leopard, Egyptian cat
	grille	convertible, grille, pickup, beach wagon, fire engine
	mushroom	agaric, mushroom, jelly fungus, gill fungus, dead-man's-fingers
	cherry	delmatian, grape, elderberry, hfordshire bullterrier, currant
	Madagascar cat	spider monkey, spider monkey, titi, indri, howler monkey


B. Leibe

53

Computer Vision WS 15/16

AlexNet Results

RWTH AACHEN UNIVERSITY



Test image Retrieved images

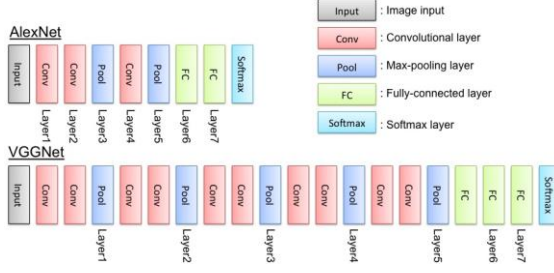
B. Leibe

54

Computer Vision WS 15/16

CNN Architectures: VGGNet (2014/15)

RWTH AACHEN UNIVERSITY



AlexNet

- Input
- Conv Layer1
- Conv Layer2
- Pool Layer3
- Conv Layer4
- Pool Layer5
- FC Layer6
- FC Layer7
- Softmax

VGGNet

- Input
- Conv Layer1
- Conv Layer2
- Pool Layer3
- Conv Layer4
- Conv Layer5
- Pool Layer6
- Conv Layer7
- Conv Layer8
- Pool Layer9
- Conv Layer10
- Conv Layer11
- FC Layer12
- FC Layer13
- FC Layer14
- Softmax

Legend:

- Input : Image input
- Conv : Convolutional layer
- Pool : Max-pooling layer
- FC : Fully-connected layer
- Softmax : Softmax layer

K. Simonyan, A. Zisserman, [Very Deep Convolutional Networks for Large-Scale Image Recognition](#), ICLR 2015

B. Leibe

55

CNN Architectures: VGGNet (2014/15)

- Main ideas
 - Deeper network
 - Stacked convolutional layers with smaller filters (+ nonlinearity)
 - Detailed evaluation of all components
- Results
 - Improved ILSVRC top-5 error rate to 6.7%.

ConvNet Configuration				
A	A+LRN	B	C	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	19 weight layers
input (224 x 224 RGB image)				
conv3-64	conv3-64 LRN	conv3-64	conv3-64	conv3-64
conv3-128	conv3-128	conv3-128	conv3-128	conv3-128
conv3-256	conv3-256	conv3-256	conv3-256	conv3-256
conv3-512	conv3-512	conv3-512	conv3-512	conv3-512
conv3-512	conv3-512	conv3-512	conv3-512	conv3-512
maxpool				
conv3-512	conv3-512	conv3-512	conv3-512	conv3-512
conv3-512	conv3-512	conv3-512	conv3-512	conv3-512
maxpool				
conv3-512	conv3-512	conv3-512	conv3-512	conv3-512
conv3-512	conv3-512	conv3-512	conv3-512	conv3-512
maxpool				
FC-4096	FC-4096	FC-4096	FC-4096	FC-4096
FC-1000	FC-1000	FC-1000	FC-1000	FC-1000
soft-max	soft-max	soft-max	soft-max	soft-max

Mainly used

B. Leibe 56

Comparison: AlexNet vs. VGGNet

- Receptive fields in the first layer
 - AlexNet: 11x11, stride 4
 - Zeiler & Fergus: 7x7, stride 2
 - VGGNet: 3x3, stride 1
- Why that?
 - If you stack three 3x3 on top of another 3x3 layer, you effectively get a 5x5 receptive field.
 - With three 3x3 layers, the receptive field is already 7x7.
 - But much fewer parameters: $3 \cdot 3^2 = 27$ instead of $7^2 = 49$.
 - In addition, non-linearities in-between 3x3 layers for additional discriminativity.

B. Leibe 57

CNN Architectures: GoogLeNet (2014)

- Main ideas
 - "Inception" module as modular component
 - Learns filters at several scales within each module

C. Szegedy, W. Liu, Y. Jia, et al, Going Deeper with Convolutions, arXiv:1409.4842, 2014.

B. Leibe 58

GoogLeNet Visualization

B. Leibe 59

Results on ILSVRC

Method	top-1 val. error (%)	top-5 val. error (%)	top-5 test error (%)
VGG (2 nets, multi-crop & dense eval.)	23.7	6.8	6.8
VGG (1 net, multi-crop & dense eval.)	24.4	7.1	7.0
VGG (ILSVRC submission, 7 nets, dense eval.)	24.7	7.5	7.3
GoogLeNet (Szegedy et al., 2014) (1 net)	-	-	7.9
GoogLeNet (Szegedy et al., 2014) (7 nets)	-	-	6.7
MSRA (He et al., 2014) (11 nets)	27.9	9.1	8.1
MSRA (He et al., 2014) (1 net)	-	-	9.1
Clarifai (Russakovsky et al., 2014) (multiple nets)	-	-	11.7
Clarifai (Russakovsky et al., 2014) (1 net)	-	-	12.5
Zeiler & Fergus (Zeiler & Fergus, 2013) (6 nets)	36.0	14.7	14.8
Zeiler & Fergus (Zeiler & Fergus, 2013) (1 net)	37.5	16.0	16.1
OverFeat (Sermanet et al., 2014) (7 nets)	34.0	13.2	13.6
OverFeat (Sermanet et al., 2014) (1 net)	35.7	14.2	-
Krizhevsky et al. (Krizhevsky et al., 2012) (5 nets)	38.1	16.4	16.4
Krizhevsky et al. (Krizhevsky et al., 2012) (1 net)	40.7	18.2	-

- VGGNet and GoogLeNet perform at similar level
 - Comparison: human performance ~5% [Karpathy]

<http://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet/>

B. Leibe 60

Newest Development: Residual Networks

B. Leibe 61

RWTH AACHEN UNIVERSITY

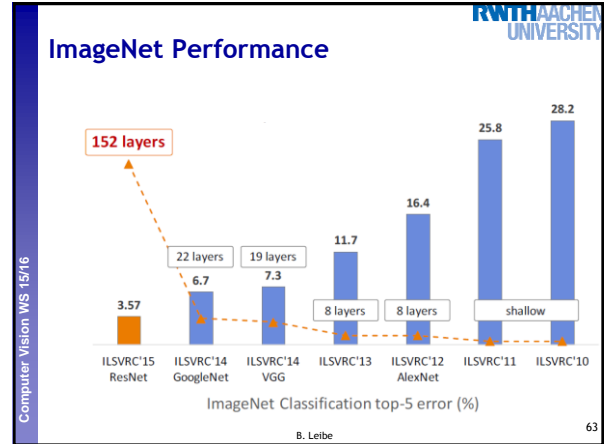
Newest Development: Residual Networks

AlexNet, 8 layers (ILSVRC 2012) VGG, 19 layers (ILSVRC 2014) ResNet, 152 layers (ILSVRC 2015)

- Core component
 - Skip connections bypassing each layer
 - Better propagation of gradients to the deeper layers

$H(x) = F(x) + x$

62

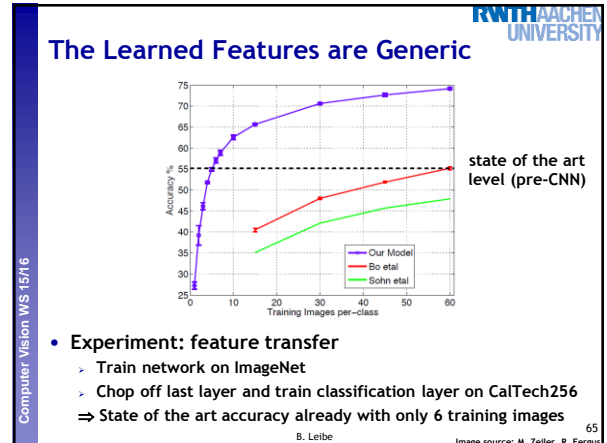


RWTH AACHEN UNIVERSITY

Topics of This Lecture

- Deep Learning
 - Motivation
- Convolutional Neural Networks
 - Convolutional Layers
 - Pooling Layers
 - Nonlinearities
- CNN Architectures
 - LeNet
 - AlexNet
 - VGGNet
 - GoogLeNet
- Applications

64



RWTH AACHEN UNIVERSITY

Other Tasks: Detection

R-CNN: Regions with CNN features

- Input image
- Extract region proposals (~2k)
- Compute CNN features
- Classify regions

Examples of classification results: aeroplane? no., person? yes., tvmonitor? no.

- Results on PASCAL VOC Detection benchmark
 - Pre-CNN state of the art: 35.1% mAP [Uijlings et al., 2013]
 - 33.4% mAP DPM
 - R-CNN: 53.7% mAP

R. Girshick, J. Donahue, T. Darrell, and J. Malik, [Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation](#), CVPR 2014

66

RWTH AACHEN UNIVERSITY

Faster R-CNN (based on ResNets)

K. He, X. Zhang, S. Ren, J. Sun, [Deep Residual Learning for Image Recognition](#), CVPR 2016.

67

RWTH AACHEN UNIVERSITY

Faster R-CNN (based on ResNets)

Computer Vision WS 15/16

K. He, X. Zhang, S. Ren, J. Sun, [Deep Residual Learning for Image Recognition](#), CVPR 2016.

B. Leibe 68

RWTH AACHEN UNIVERSITY

Other Tasks: Semantic Segmentation

Computer Vision WS 15/16

[Farabet et al. ICML 2012, PAMI 2013]

B. Leibe 69

RWTH AACHEN UNIVERSITY

Semantic Segmentation

Computer Vision WS 15/16

[Pohlen, Hermans, Mathias, Leibe, arXiv 2016]

- More recent results
 - Based on an extension of ResNets

B. Leibe 70

RWTH AACHEN UNIVERSITY

Other Tasks: Face Verification

Computer Vision WS 15/16

Y. Taigman, M. Yang, M. Ranzato, L. Wolf, [DeepFace: Closing the Gap to Human-Level Performance in Face Verification](#), CVPR 2014

Slide credit: Svetlana Lazebnik

72

RWTH AACHEN UNIVERSITY

Commercial Recognition Services

- E.g., **clarifai**

Computer Vision WS 15/16

- Be careful when taking test images from Google Search
 - Chances are they may have been seen in the training set...

B. Leibe 73

RWTH AACHEN UNIVERSITY

Commercial Recognition Services

Computer Vision WS 15/16

B. Leibe 74

References and Further Reading

- **LeNet**
 - Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, [Gradient-based learning applied to document recognition](#), Proceedings of the IEEE 86(11): 2278-2324, 1998.
- **AlexNet**
 - A. Krizhevsky, I. Sutskever, and G. Hinton, [ImageNet Classification with Deep Convolutional Neural Networks](#), NIPS 2012.
- **VGGNet**
 - K. Simonyan, A. Zisserman, [Very Deep Convolutional Networks for Large-Scale Image Recognition](#), ICLR 2015
- **GoogLeNet**
 - C. Szegedy, W. Liu, Y. Jia, et al, [Going Deeper with Convolutions](#), arXiv:1409.4842, 2014.