



Machine Learning Winter '19

## Recap: GoogLeNet (2014)

RWTH AACHEN UNIVERSITY

- Ideas:
  - Learn features at multiple scales
  - Modular structure

(b) Inception module with dimension reductions

B. Leibe

Image source: Szepesvári et al.

8

Machine Learning Winter '19

## Discussion

RWTH AACHEN UNIVERSITY

- GoogLeNet
  - 12x fewer parameters than AlexNet
  - ~5M parameters
  - Where does the main reduction come from?
    - From throwing away the fully connected (FC) layers.
- Effect
  - After last pooling layer, volume is of size  $[7 \times 7 \times 1024]$
  - Normally you would place the first 4096-D FC layer here (Many million params).
  - Instead: use Average pooling in each depth slice:
    - Reduces the output to  $[1 \times 1 \times 1024]$ .
  - Performance actually improves by 0.6% compared to when using FC layers (less overfitting?)

Slide credit: Andrej Karpathy

B. Leibe

Image source: Szepesvári et al.

9

Machine Learning Winter '19

## Topics of This Lecture

RWTH AACHEN UNIVERSITY

- Recap: CNN Architectures
- Residual Networks
  - Detailed analysis
  - ResNets as ensembles of shallow networks
- Visualizing CNNs
  - Visualizing CNN features
  - Visualizing responses
  - Visualizing learned structures
- Applications of CNNs
  - Object detection
  - Semantic segmentation
  - Face identification

B. Leibe

11

Machine Learning Winter '19

## Recap: Residual Networks

RWTH AACHEN UNIVERSITY

AlexNet, 8 layers (ILSVRC 2012)

VGG, 19 layers (ILSVRC 2014)

GoogLeNet, 22 layers (ILSVRC 2014)

Slide credit: Kaiming He

B. Leibe

12

Machine Learning Winter '19

## Recap: Residual Networks

RWTH AACHEN UNIVERSITY

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

B. Leibe

13

Machine Learning Winter '19

## Spectrum of Depth

RWTH AACHEN UNIVERSITY

shallow ← → deeper

5 layers: easy

>10 layers: initialization, Batch Normalization

>30 layers: skip connections

>100 layers: identity skip connections

>1000 layers: ?

Slide credit: Kaiming He

B. Leibe

14

### Spectrum of Depth

- Deeper models are more powerful
  - But training them is harder.
  - Main problem: getting the gradients back to the early layers
  - The deeper the network, the more effort is required for this.

Machine Learning Winter '19 | Slide adapted from Kaiming He | B. Leibe | 15

### Initialization

22-layer ReLU net:  
good init converges faster

30-layer ReLU net:  
good init is able to converge

- Importance of proper initialization (Recall Lecture 15)
  - Glorot initialization for tanh nonlinearities
  - He initialization for ReLU nonlinearities
  - ⇒ For deep networks, this really makes a difference!

Machine Learning Winter '19 | Slide credit: Kaiming He | B. Leibe | 16

### Batch Normalization

- Effect of batch normalization
  - Greatly improved speed of convergence
  - Often better accuracy achievable

Machine Learning Winter '19 | B. Leibe | Image source: Ioffe & Szegedy | 17

### Going Deeper

- Checklist
  - Initialization ok
  - Batch normalization ok
  - Are we now set?
    - Is learning better networks now as simple as stacking more layers?

Machine Learning Winter '19 | Slide credit: Kaiming He | B. Leibe | 18

### Simply Stacking Layers?

CIFAR-10  
train error (%)

CIFAR-10  
test error (%)

- Experiment going deeper
  - Plain nets: stacking 3x3 convolution layers
  - ⇒ 56-layer net has higher training error than 20-layer net

Machine Learning Winter '19 | Slide credit: Kaiming He | B. Leibe | 19

### Simply Stacking Layers?

CIFAR-10

ImageNet-1000

- General observation
  - Overly deep networks have higher training error
  - A general phenomenon, observed in many training sets

Machine Learning Winter '19 | Slide credit: Kaiming He | B. Leibe | 20

### Why Is That???

- A deeper model should not have higher training error!
  - Richer solution space should allow it to find better solutions
- Solution by construction
  - Copy the original layers from a learned shallower model
  - Set the extra layers as identity
  - Such a network should achieve at least the same low training error.
- Reason: Optimization difficulties
  - Solvers cannot find the solution when going deeper...

Machine Learning Winter '19 | RWTH AACHEN UNIVERSITY | Slide credit: Kaiqing He | B. Leibe | 21

### Deep Residual Learning

- Plain net
  - any two stacked layers
  - weight layer
  - relu
  - weight layer
  - relu
  - $H(x)$
- $H(x)$  is any desired mapping
- Hope the 2 weight layers fit  $H(x)$

Machine Learning Winter '19 | RWTH AACHEN UNIVERSITY | Slide credit: Kaiqing He | B. Leibe | 22

### Deep Residual Learning

- Residual net
  - $F(x)$
  - weight layer
  - relu
  - weight layer
  - relu
  - $H(x) = F(x) + x$
  - identity
  - $x$
- $H(x)$  is any desired mapping
- Hope the 2 weight layers fit  $H(x)$
- Hope the 2 weight layers fit  $F(x)$
- Let  $H(x) = F(x) + x$

Machine Learning Winter '19 | RWTH AACHEN UNIVERSITY | Slide credit: Kaiqing He | B. Leibe | 23

### Deep Residual Learning

- $F(x)$  is a residual mapping w.r.t. identity
  - If identity were optimal, it is easy to set weights as 0
  - If optimal mapping is closer to identity, it is easier to find small fluctuations
  - Further advantage: direct path for the gradient to flow to the previous stages

Machine Learning Winter '19 | RWTH AACHEN UNIVERSITY | Slide credit: Kaiqing He | B. Leibe | 24

### Network Design

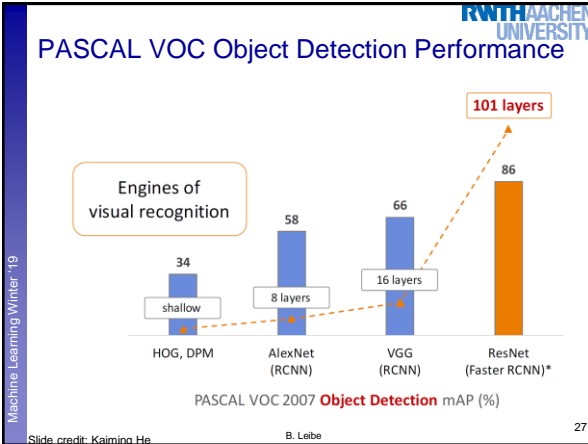
- Simple, VGG-style design
  - (Almost) all 3x3 convolutions
  - Spatial size / 2  $\Rightarrow$  #filters  $\cdot$  2 (same complexity per layer)
  - Batch normalization
  - $\Rightarrow$  Simple design, just deep.

Machine Learning Winter '19 | RWTH AACHEN UNIVERSITY | Slide credit: Kaiqing He | B. Leibe | 25

### ImageNet Performance

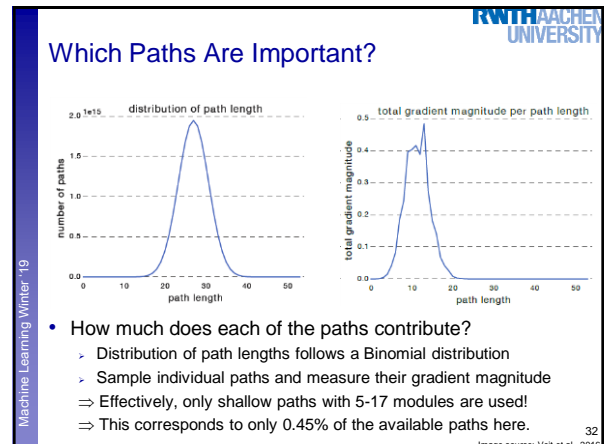
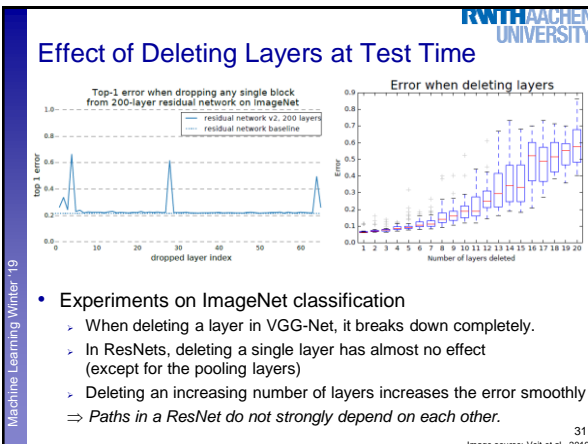
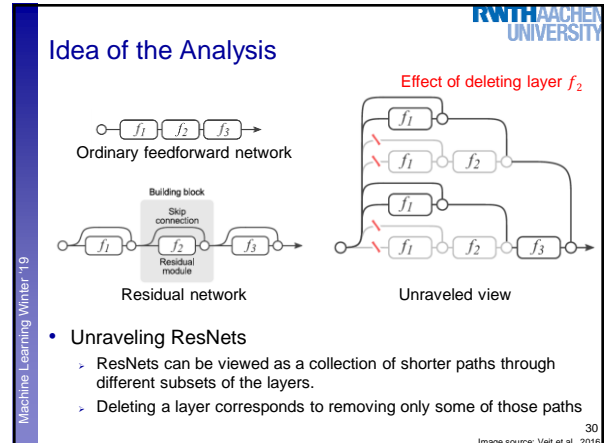
Model	Layers	ImageNet Classification top-5 error (%)
ILSVRC'15 ResNet	152	3.57
ILSVRC'14 GoogleNet	22	6.7
ILSVRC'14 VGG	19	7.3
ILSVRC'13 AlexNet	8	11.7
ILSVRC'12 AlexNet	8	16.4
ILSVRC'11 shallow	shallow	25.8
ILSVRC'10	shallow	28.2

Machine Learning Winter '19 | RWTH AACHEN UNIVERSITY | Slide credit: Kaiqing He | B. Leibe | 26



- ### Topics of This Lecture
- Recap: CNN Architectures
  - Residual Networks
    - Detailed analysis
    - ResNets as ensembles of shallow networks
  - Applications of CNNs
    - Object detection
    - Semantic segmentation
    - Face identification

- ### What Is The Secret Behind ResNets?
- Empirically, they perform very well, but why is that?
  - He's original explanation [He, 2016]
    - ResNets allow gradients to pass through the skip connections in unchanged form.
    - This makes it possible to effectively train deeper networks.
    - ⇒ Secret of success: **depth is good**
  - More recent explanation [Veit, 2016]
    - ResNets actually do not use deep network paths.
    - Instead, they effectively implement an ensemble of shallow network paths.
    - ⇒ Secret of success: **ensembles are good**
- A. Veit, M. Wilber, S. Belongie, *Residual Networks Behave Like Ensembles of Relatively Shallow Networks*, NIPS 2016



Machine Learning Winter '19

## Summary

- The effective paths in ResNets are relatively shallow
  - Effectively only 5-17 active modules
- This explains the resilience to deletion
  - Deleting any single layer only affects a subset of paths (and the shorter ones less than the longer ones).
- New interpretation of ResNets
  - ResNets work by creating an ensemble of relatively shallow paths
  - Making ResNets deeper increases the size of this ensemble
  - Excluding longer paths from training does not negatively affect the results.

Building block: skip connection, Residual module,  $f_1$ ,  $f_2$

total gradient magnitude per path length

path length

33  
Image source: Vait et al., 2014

Machine Learning Winter '19

## Topics of This Lecture

- Recap: CNN Architectures
- Residual Networks
  - Detailed analysis
  - ResNets as ensembles of shallow networks
- Visualizing CNNs
  - Visualizing CNN features
  - Visualizing responses
  - Visualizing learned structures
- Applications of CNNs
  - Object detection
  - Semantic segmentation
  - Face identification

B. Leibe

34

Machine Learning Winter '19

## Visualizing CNNs

DeconvNet

ConvNet

Layer Above Reconstruction

Max Unpooling

Unpooled Maps

Rectified Linear Function

Rectified Unpooled Maps

Convolutional Filtering (F)

Reconstruction

Switches

Pooled Maps

Max Pooling

Rectified Feature Maps

Rectified Linear Function

Feature Maps

Convolutional Filtering (F)

Layer Below Pooled Maps

Unpooling

Max Locations "Switches"

Rectified Feature Maps

Pooling

35  
Image source: M. Zeiler, R. Fergus

Machine Learning Winter '19

## Visualizing CNNs

Layer 1

Layer 2

reconstruction of image patches from that unit (indicates aspect of patches which unit is sensitive to)

top 9 image patches that cause maximal activation in layer 2 unit

M. Zeiler, R. Fergus, [Visualizing and Understanding Convolutional Neural Networks](#), ECCV 2014.

Slide credit: Richard Turner

B. Leibe

36  
Image source: M. Zeiler, R. Fergus

Machine Learning Winter '19

## Visualizing CNNs

Layer 3

B. Leibe

37  
Image source: M. Zeiler, R. Fergus

Machine Learning Winter '19

## Visualizing CNNs

Layer 4

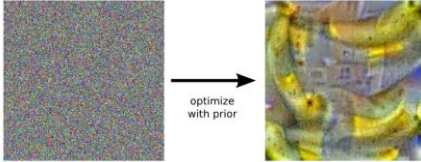
Layer 5

B. Leibe

38  
Image source: M. Zeiler, R. Fergus

Machine Learning Winter '19

## Inceptionism: Dreaming ConvNets



- Idea
  - Start with a random noise image.
  - Enhance the input image such as to enforce a particular response (e.g., banana).
  - Combine with prior constraint that image should have similar statistics as natural images.

⇒ Network hallucinates characteristics of the learned class.

<http://googleresearch.blogspot.de/2015/06/inceptionism-going-deeper-into-neural.html>

39

Machine Learning Winter '19

## Inceptionism: Dreaming ConvNets

- Results



<http://googleresearch.blogspot.de/2015/07/deepdream-code-example-for-visualizing.html>

40

Machine Learning Winter '19

## Inceptionism: Dreaming ConvNets



<https://www.youtube.com/watch?v=IREsx-xW00g>

41

Machine Learning Winter '19

## Topics of This Lecture

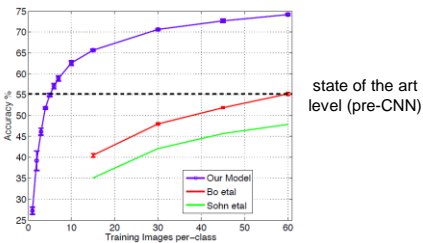
- Recap: CNN Architectures
- Residual Networks
  - Detailed analysis
  - ResNets as ensembles of shallow networks
- Visualizing CNNs
  - Visualizing CNN features
  - Visualizing responses
  - Visualizing learned structures
- Applications of CNNs
  - Object detection
  - Semantic segmentation
  - Face identification

B. Leibe

42

Machine Learning Winter '19

## The Learned Features are Generic



- Experiment: feature transfer
  - Train AlexNet-like network on ImageNet
  - Chop off last layer and train classification layer on CalTech256

⇒ State of the art accuracy already with only 6 training images!

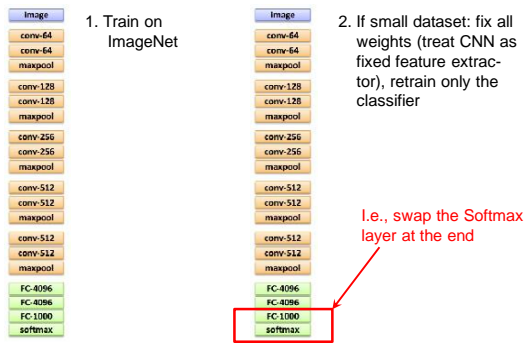
B. Leibe

43

Machine Learning Winter '19

## Transfer Learning with CNNs

1. Train on ImageNet
2. If small dataset: fix all weights (treat CNN as fixed feature extractor), retrain only the classifier



I.e., swap the Softmax layer at the end

Slide credit: Andrej Karpathy

B. Leibe

44

## Transfer Learning with CNNs

1. Train on ImageNet

3. If you have medium sized dataset, "finetune" instead: use the old weights as initialization, train the full network or only some of the higher layers.

Slide credit: Andrei Karpathy. B. Leibe. 45

## Other Tasks: Detection

### R-CNN: Regions with CNN features

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

- 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. 48

## More Recent Version: Faster R-CNN

- One network, four losses
  - Remove dependence on external region proposal algorithm.
  - Instead, infer region proposals from same CNN.
  - Feature sharing
  - Joint training

⇒ Object detection in a single pass becomes possible.

Slide credit: Ross Girshick. 49

## Faster R-CNN (based on ResNets)

K. He, X. Zhang, S. Ren, J. Sun, [Deep Residual Learning for Image Recognition](#), CVPR 2016. B. Leibe. 50

## Faster R-CNN (based on ResNets)

K. He, X. Zhang, S. Ren, J. Sun, [Deep Residual Learning for Image Recognition](#), CVPR 2016. B. Leibe. 51

## Most Recent Version: Mask R-CNN

K. He, G. Gkioxari, P. Dollár, R. Girshick, [Mask R-CNN](#), arXiv 1703.06870. 52



**Mask R-CNN Results**

- Detection + Instance segmentation
- Detection + Pose estimation

Machine Learning Winter '19

Figure credit: K. He, G. Gkioxari, P. Dollár, R. Girshick

**YOLO / SSD**

Input image  $3 \times H \times W$

Divide image into grid  $7 \times 7$

- Idea: Directly go from image to detection scores
- Within each grid cell
  - Start from a set of anchor boxes
  - Regress from each of the B anchor boxes to a final box
  - Predict scores for each of C classes (including background)

Machine Learning Winter '19

Slide credit: FeiFei Li

**YOLO**

J. Redmon, S. Divvala, R. Girshick, A. Farhadi, [You Only Look Once: Unified, Real-Time Object Detection](#), CVPR 2016.

Machine Learning Winter '19

**Object Detection Performance**

PASCAL VOC

mean Average Precision (mAP)

year

Before deep convnets

Using deep convnets

RCNN

Fast RCNN

Faster RCNN

Machine Learning Winter '19

Slide credit: Ross Girshick

B. Leibe

**Semantic Image Segmentation**

forward/inference

backward/learning

pixel-wise prediction

segmentation, g.t.

- Perform pixel-wise prediction task
  - Usually done using **Fully Convolutional Networks (FCNs)**
    - All operations formulated as convolutions
    - Advantage: can process arbitrarily sized images

Machine Learning Winter '19

Image sources: Long, Shelhamer, Darrell

**CNNs vs. FCNs**

- CNN
- FCN

“tabby cat”

convolutionalization

tabby cat heatmap

- Intuition
  - Think of FCNs as performing a sliding-window classification, producing a heatmap of output scores for each class

Machine Learning Winter '19

Image sources: Long, Shelhamer, Darrell

Machine Learning Winter '19

## Semantic Image Segmentation

- Encoder-Decoder Architecture
  - Problem: FCN output has low resolution
  - Solution: perform upsampling to get back to desired resolution
  - Use skip connections to preserve higher-resolution information

61  
Image source: Newell et al.

Machine Learning Winter '19

## Semantic Segmentation

- Current state-of-the-art
  - Based on an extension of ResNets

[Pohlen, Hermans, Mathias, Leibe, CVPR 2017]

Machine Learning Winter '19

## Other Tasks: Face Identification

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

63

Machine Learning Winter '19

## Learning Similarity Functions

- Siamese Network
  - Present the two stimuli to two identical copies of a network (with shared parameters)
  - Train them to output similar values if the inputs are (semantically) similar.
- Used for many matching tasks
  - Face identification
  - Stereo estimation
  - Optical flow
  - ...

B. Leibe

64

Machine Learning Winter '19

## Extension: Triplet Loss Networks

- Learning a discriminative embedding
  - Present the network with triplets of examples

- Apply triplet loss to learn an embedding  $f(\cdot)$  that groups the positive example closer to the anchor than the negative one.
 
$$\|f(x_i^a) - f(x_i^p)\|_2^2 < \|f(x_i^a) - f(x_i^n)\|_2^2$$

$\Rightarrow$  Used with great success in Google's FaceNet face identification

B. Leibe

65

Machine Learning Winter '19

## References and Further Reading

- ResNets
  - K. He, X. Zhang, S. Ren, J. Sun, [Deep Residual Learning for Image Recognition](#), CVPR 2016.
  - A. Veit, M. Wilber, S. Belongie, [Residual Networks Behave Like Ensembles of Relatively Shallow Networks](#), NIPS 2016.

B. Leibe

67

## References: Computer Vision Tasks

- Object Detection
  - R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, CVPR 2014.
  - S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, NIPS 2015.
  - J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You Only Look Once: Unified Real-Time Object Detection, CVPR 2016.
  - W. Liu, D. Anguelov, [D. Erhan](#), [C. Szegedy](#), S. Reed, C-Y. Fu, A.C. Berg, SSD: Single Shot Multi Box Detector, ECCV 2016.

## References: Computer Vision Tasks

- Semantic Segmentation
  - J. Long, E. Shelhamer, T. Darrell, Fully Convolutional Networks for Semantic Segmentation, CVPR 2015.
  - H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid Scene Parsing Network, arXiv 1612.01105, 2016.