RWTH AACHEN UNIVERSITY

Convolutional Neural Networks in Computer Vision

21.02.2017

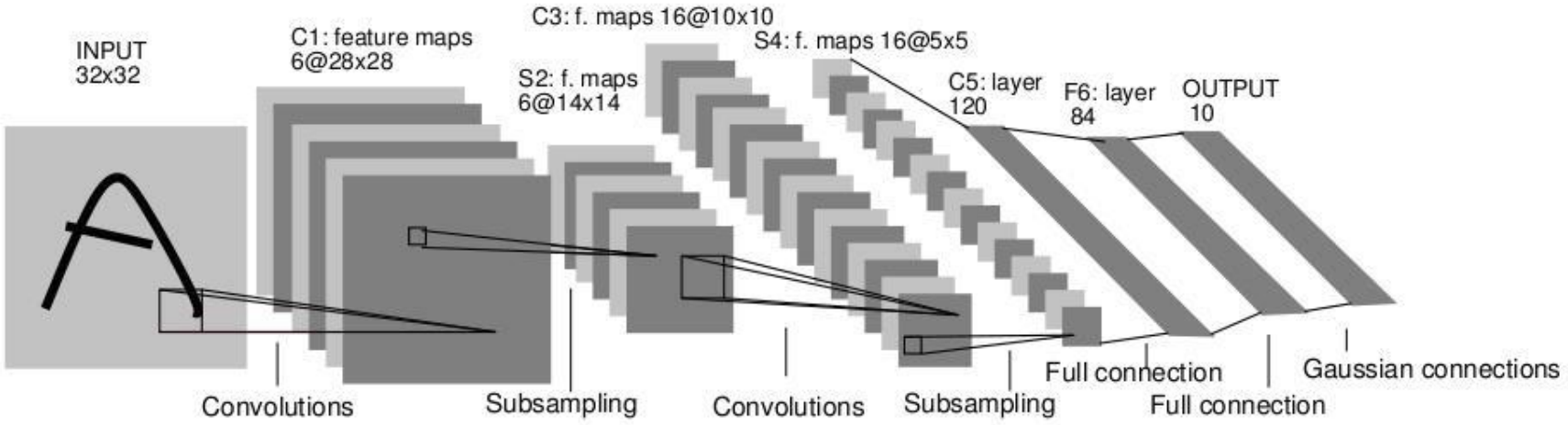
Bastian Leibe
RWTH Aachen
http://www.vision.rwth-aachen.de/

leibe@vision.rwth-aachen.de

Outline

- **Recap: CNNs**
  - Convolutional layers
  - Pooling layers
- CNN Architectures and Historical Development
  - LeNet
  - AlexNet
  - VGGNet
  - GoogLeNet
  - ResNet
  - Fully Convolutional Networks
- Applications
  - Object Detection
  - Semantic Image Segmentation
  - Matching

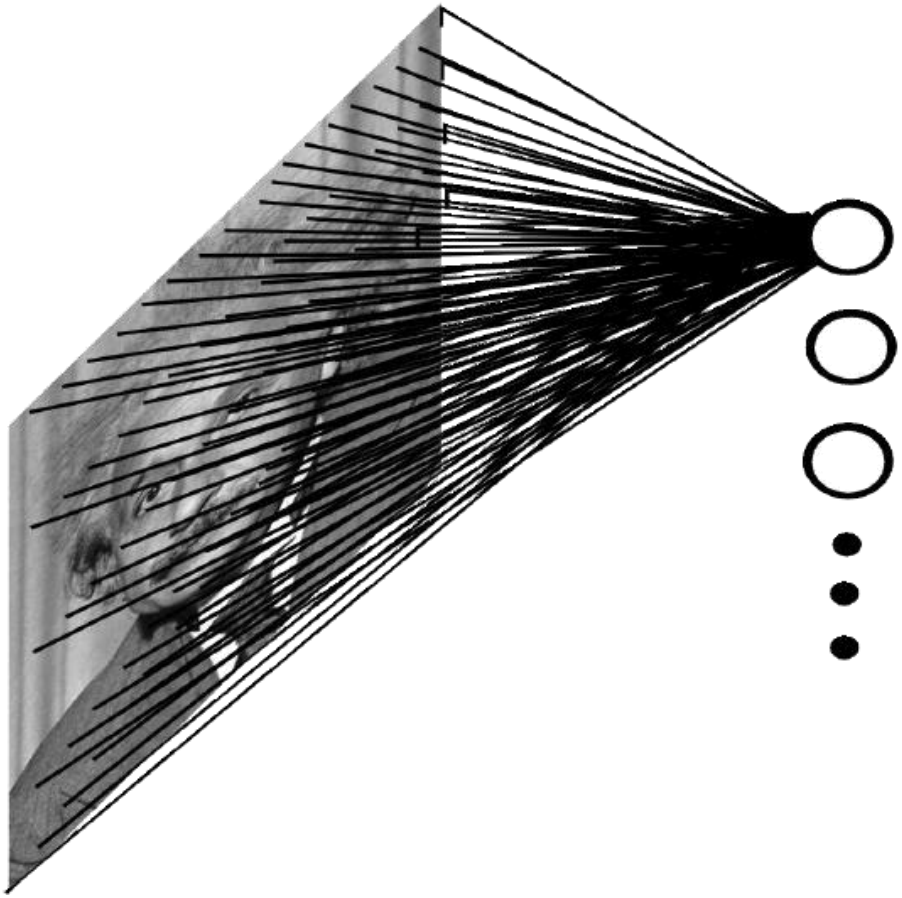
Convolutional Neural Networks (CNN, ConvNet)



- **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.

Convolutional Networks: Intuition

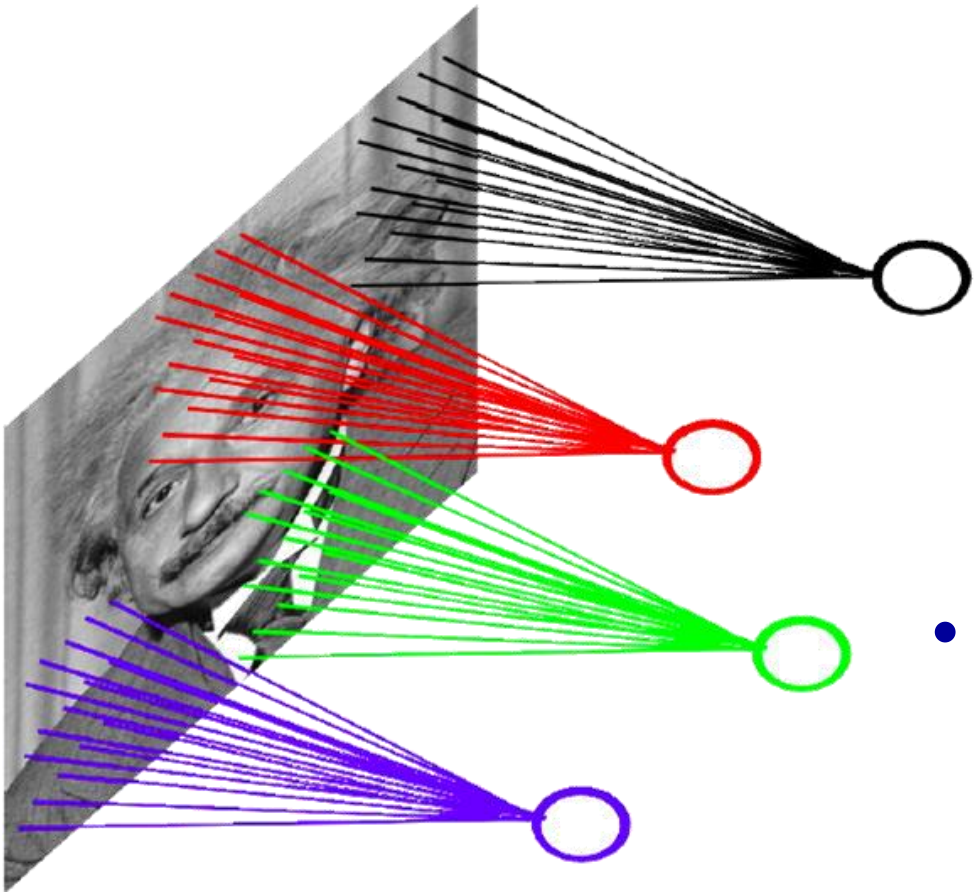


- Fully connected network
  - E.g. 1000×1000 image
    1M hidden units

  ⇒ 1T parameters!

- Ideas to improve this
  - Spatial correlation is local

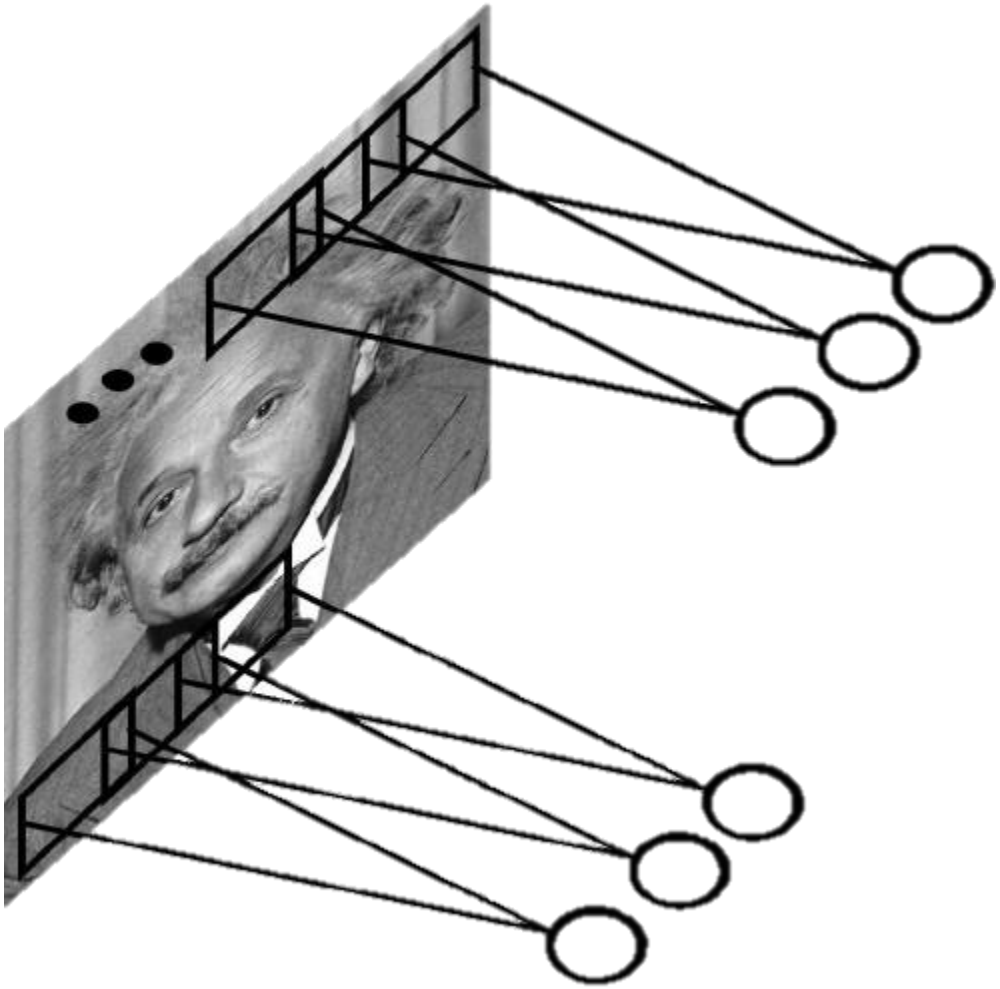
Convolutional Networks: Intuition



- 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

Convolutional Networks: Intuition

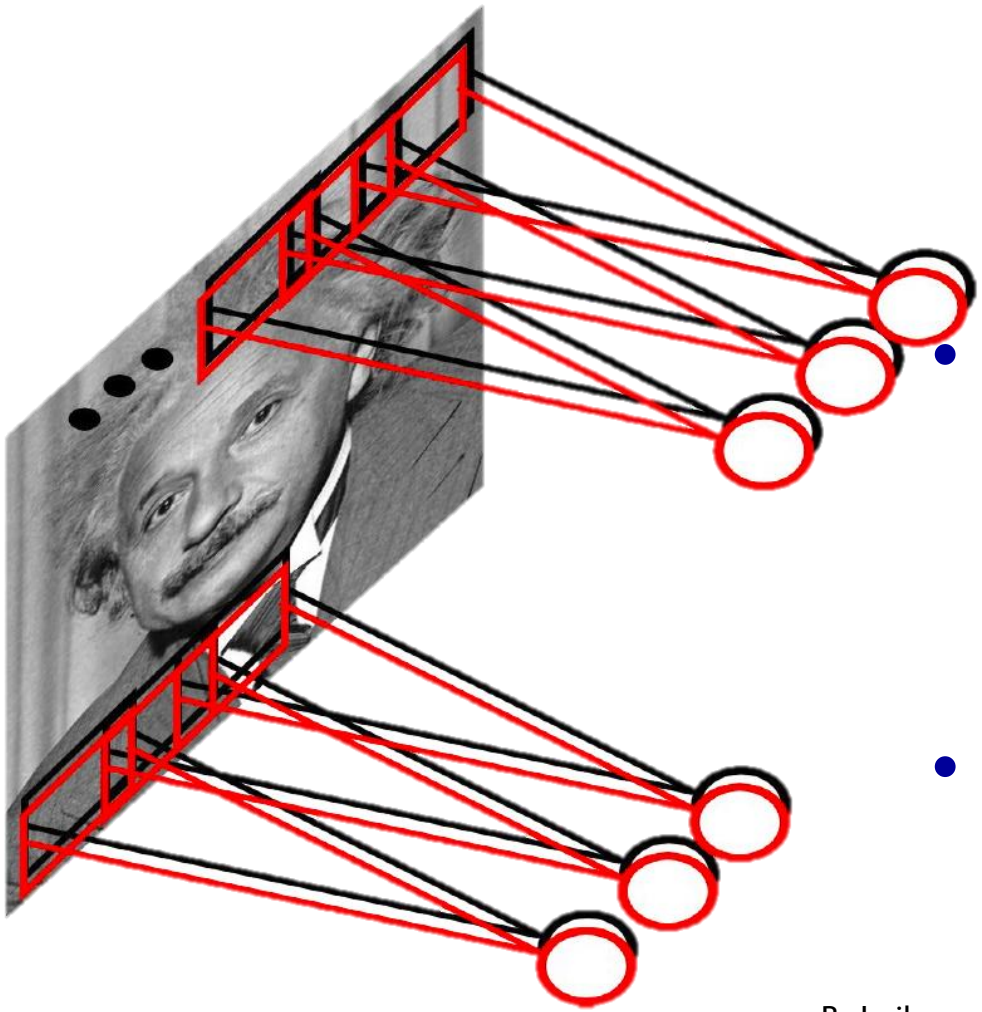


- Convolutional net
  - Share the same parameters across different locations
  - Convolutions with learned kernels

Slide adapted from Marc'Aurelio Ranzato

Image source: Yann LeCun

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

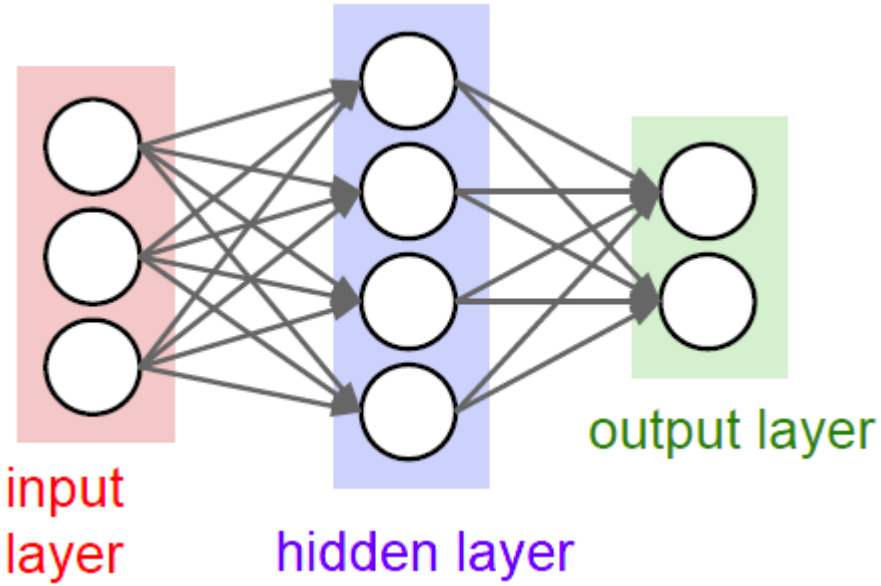
  ⇒ 10k parameters
- **Result: Response map**
  - size: 1000×1000×100
  - Only memory, not params!

Slide adapted from Marc'Aurelio Ranzato

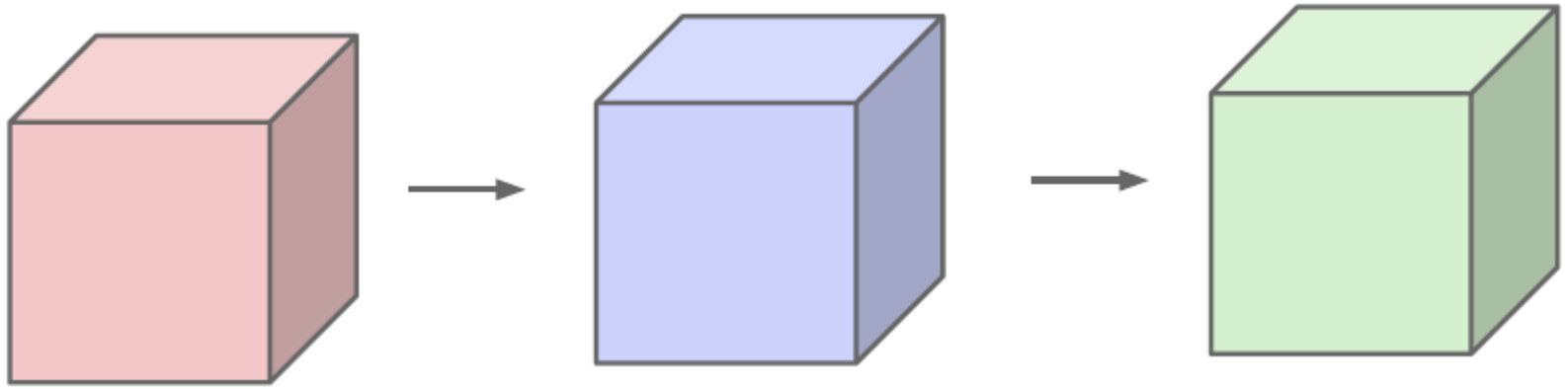
Image source: Yann LeCun

Important Conceptual Shift

- **Before**

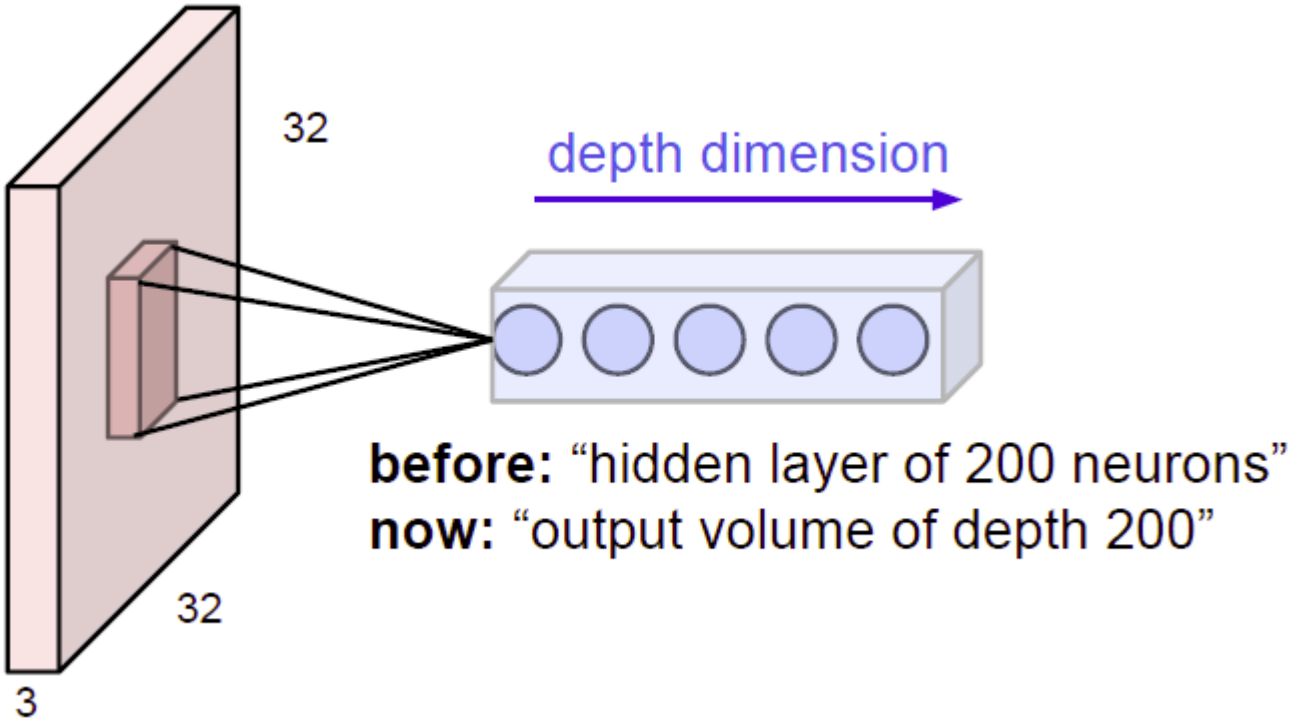


- **Now:**



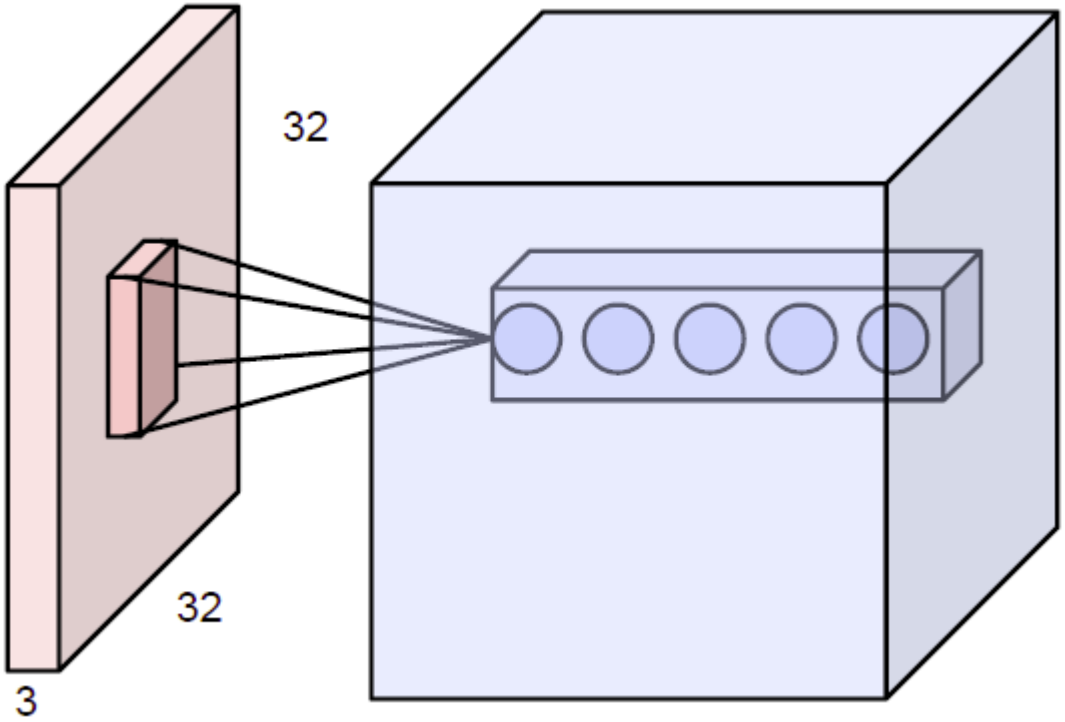
Slide credit: FeiFei Li, Andrej Karpathy

Convolution Layers

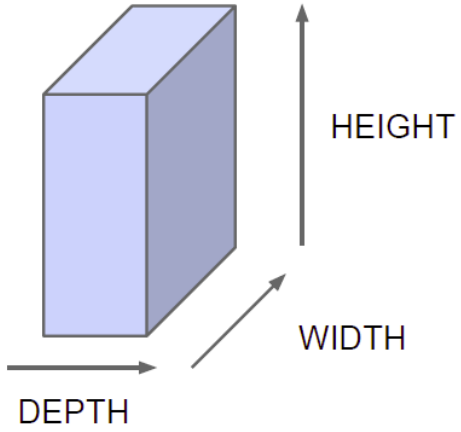


- All Neural Net activations arranged in 3 dimensions
  - Multiple neurons all looking at the same input region, stacked in depth

Convolution Layers

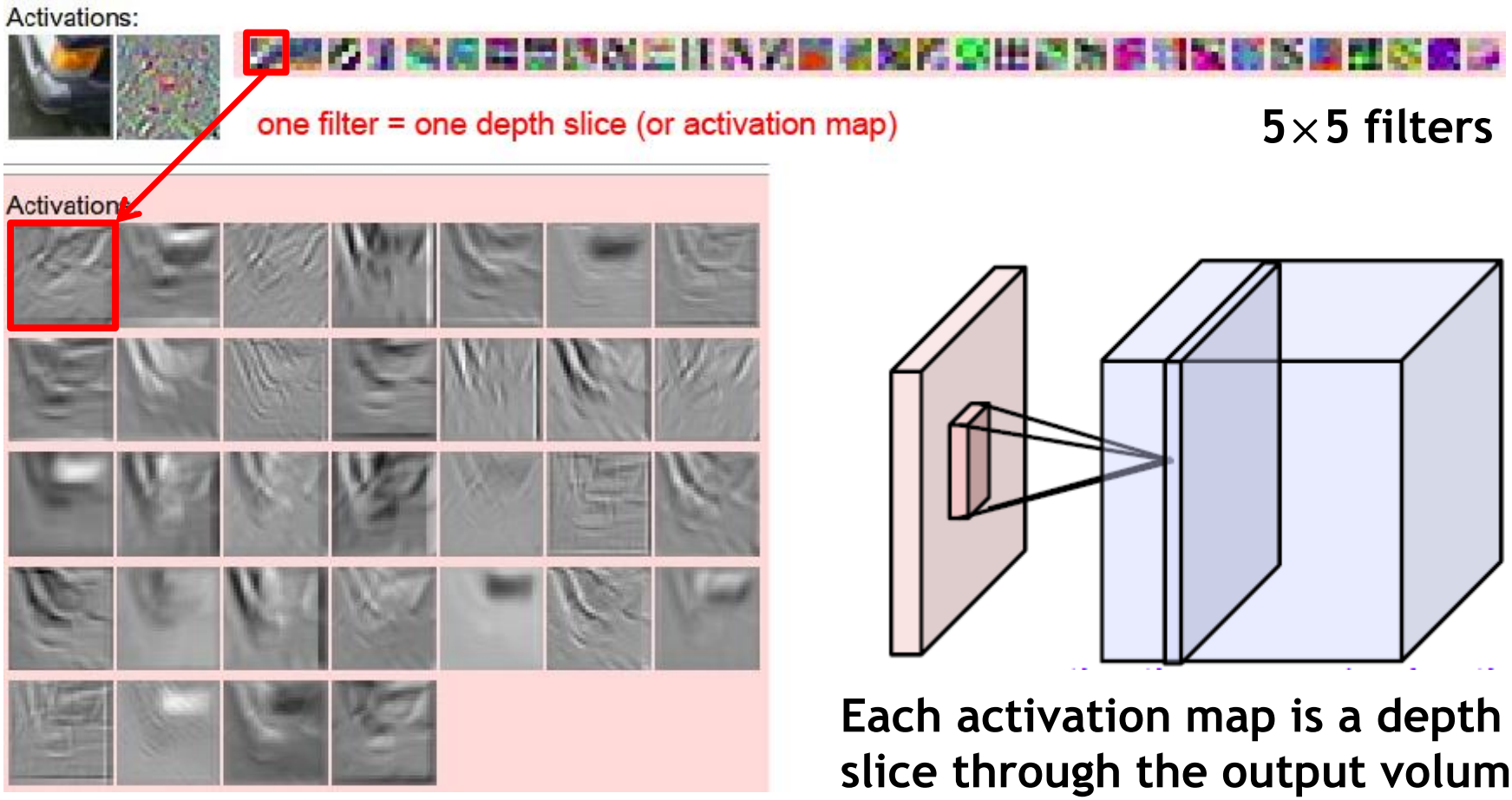


Naming convention:

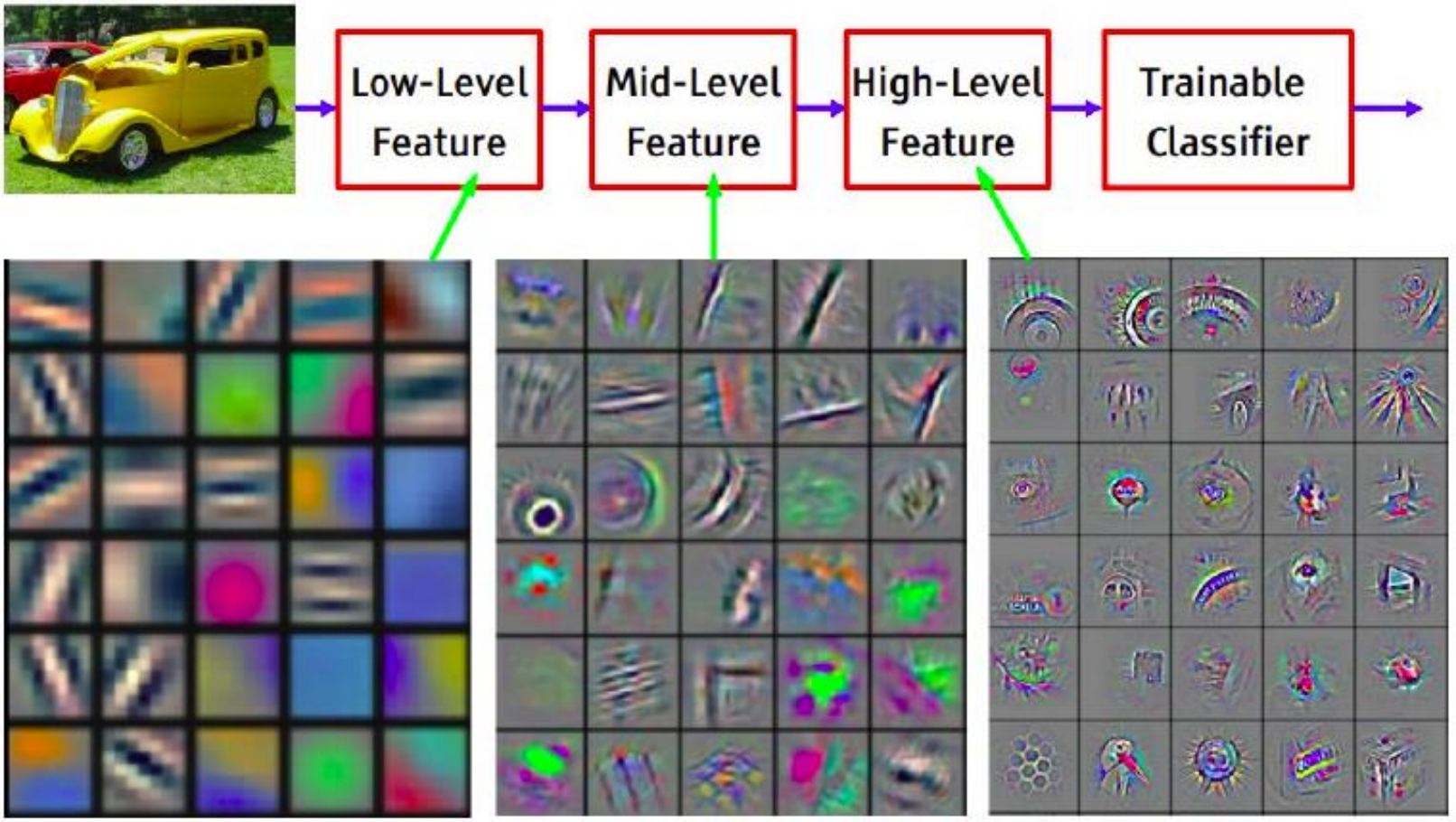


- All Neural Net activations arranged in 3 dimensions
  - Multiple neurons all looking at the same input region, stacked in depth
  - Form a single [1×1×depth] depth column in output volume.

Activation Maps of Convolutional Filters

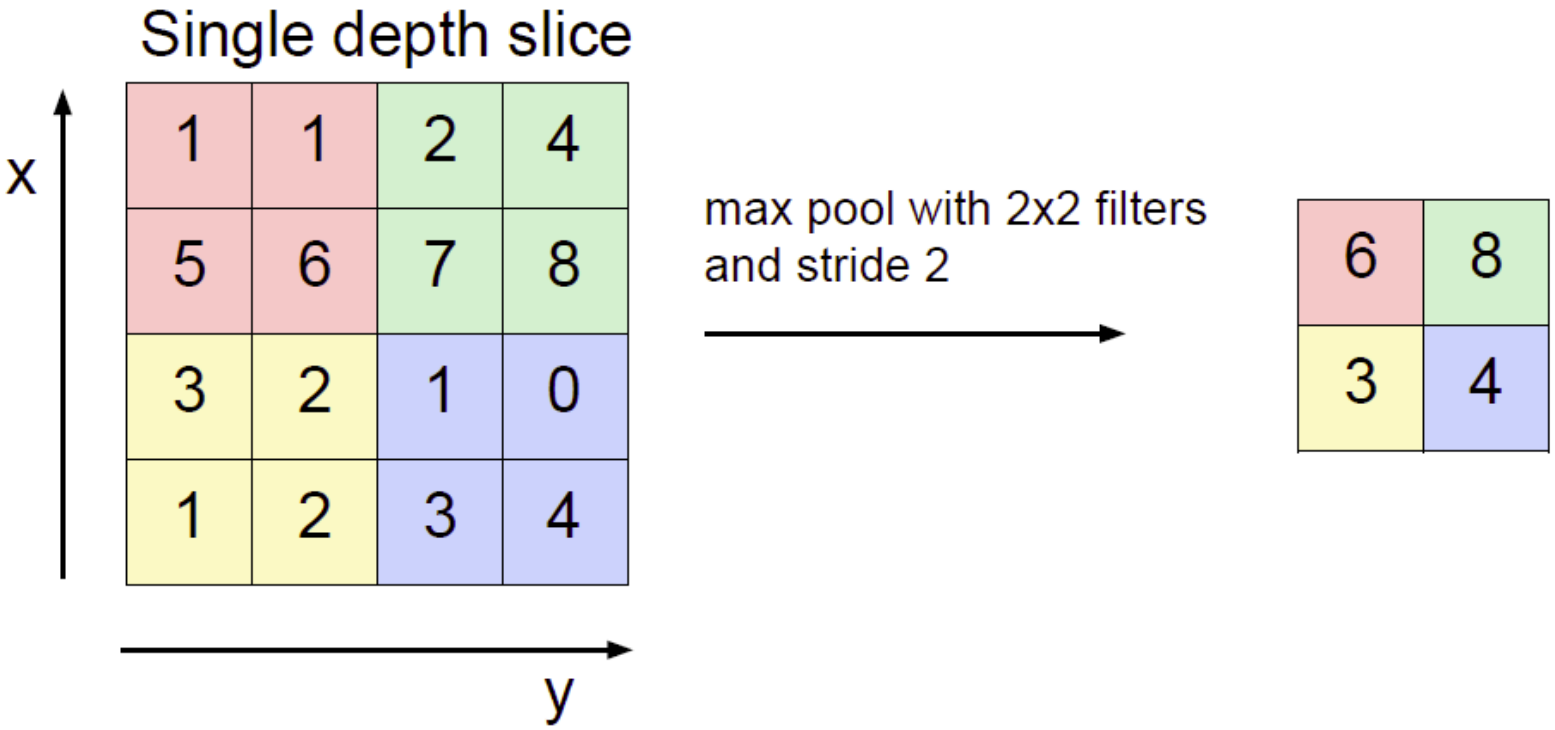


Effect of Multiple Convolution Layers



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

Max Pooling

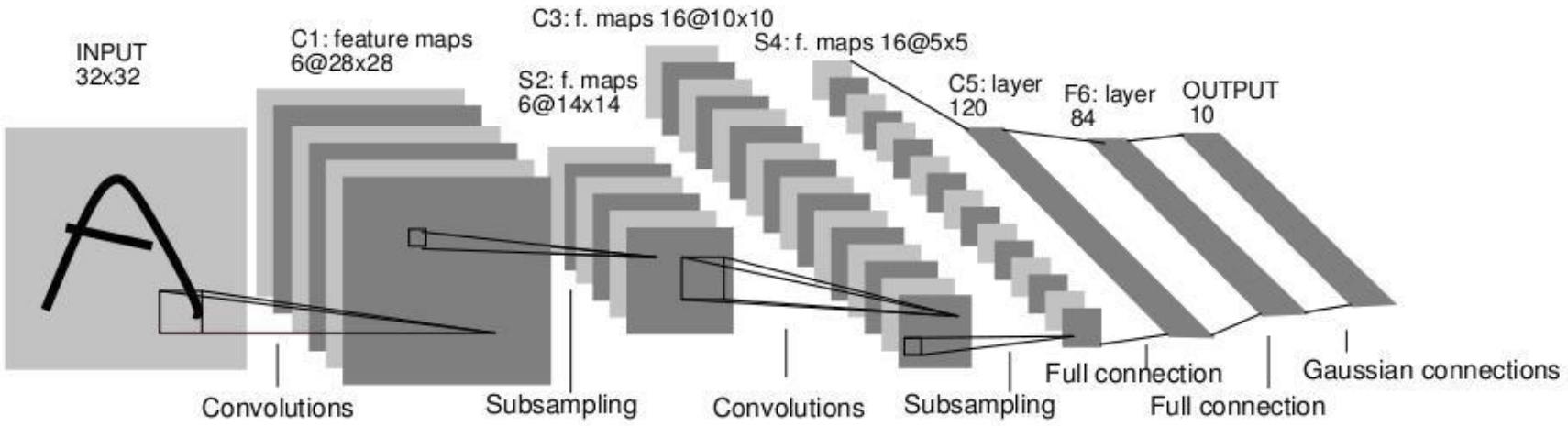


- **Effect:**
  - **Make the representation smaller without losing too much information**
  - **Achieve robustness to translations**

Outline

- Recap: CNNs
  - Convolutional layers
  - Pooling layers
- **CNN Architectures and Historical Development**
  - **LeNet**
  - **AlexNet**
  - **VGGNet**
  - **GoogLeNet**
  - **ResNet**
  - **Fully Convolutional Networks**
- Applications
  - Object Detection
  - Semantic Image Segmentation
  - Matching

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)**
  - **Difficulties scaling this to larger images (beyond 32×32)**

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.

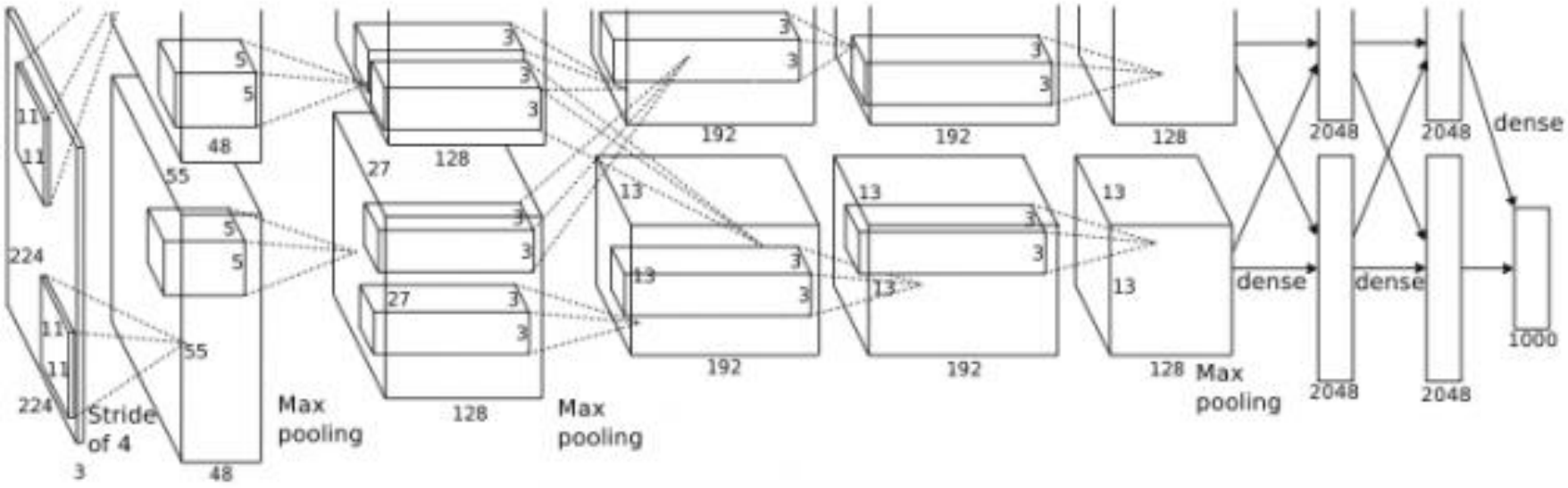
ImageNet Challenge 2012

- ImageNet
  - ~14M labeled internet images
  - 20k classes
  - Human labels via Amazon Mechanical Turk

[Deng et al., CVPR'09]

- Challenge (ILSVRC)
  - 1.2 million training images
  - 1000 classes
  - Goal: Predict ground-truth class within top-5 responses
  - Human level performance ~5%
  - Currently one of the top benchmarks in Computer Vision

CNN Architectures: AlexNet (2012)

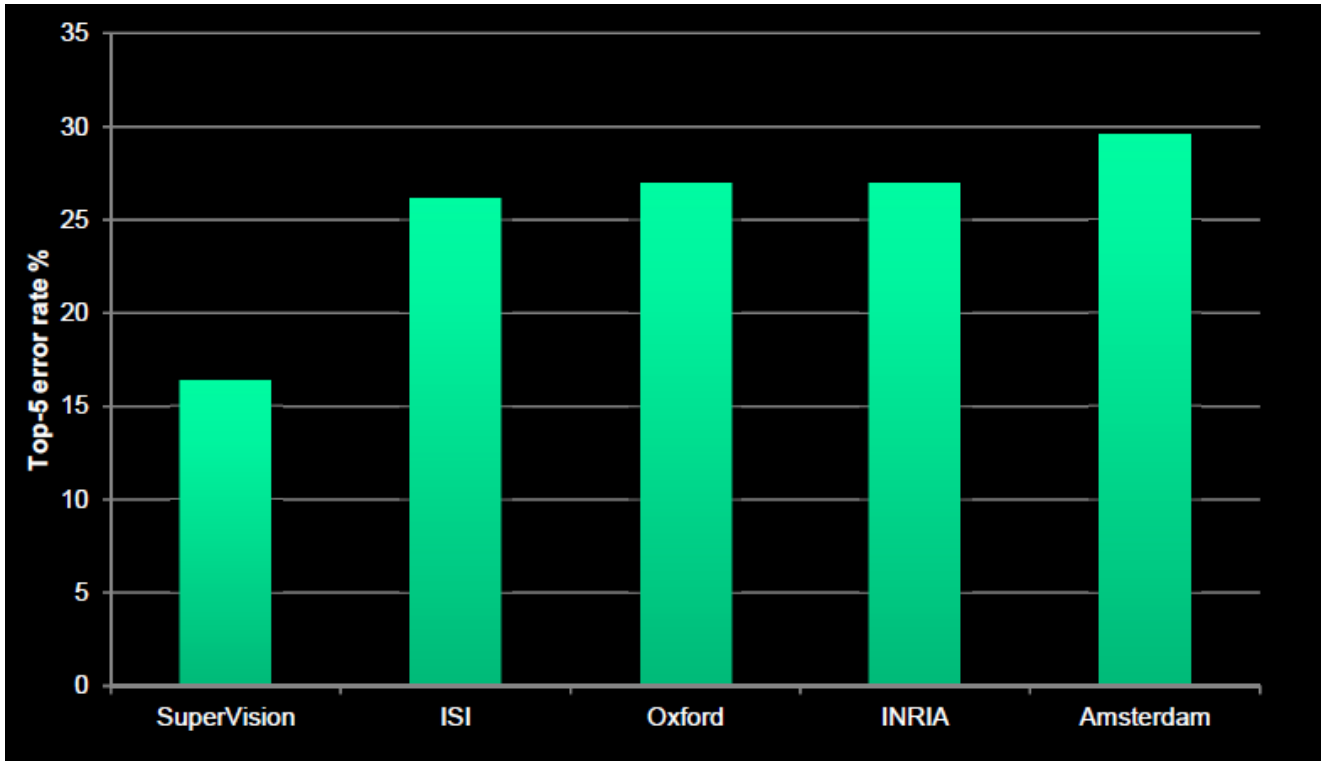


- **Similar framework as LeNet, but**
  - **Bigger model (7 hidden layers, 650k units, 60M parameters)**
  - **More data (10^6 images instead of 10^3)**
  - **GPU implementation**
  - **Better regularization and up-to-date tricks for training (Dropout)**

A. Krizhevsky, I. Sutskever, and G. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, NIPS 2012.

Image source: A. Krizhevsky, I. Sutskever and G.E. Hinton, NIPS 2012

ILSVRC 2012 Results



- 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

AlexNet Results

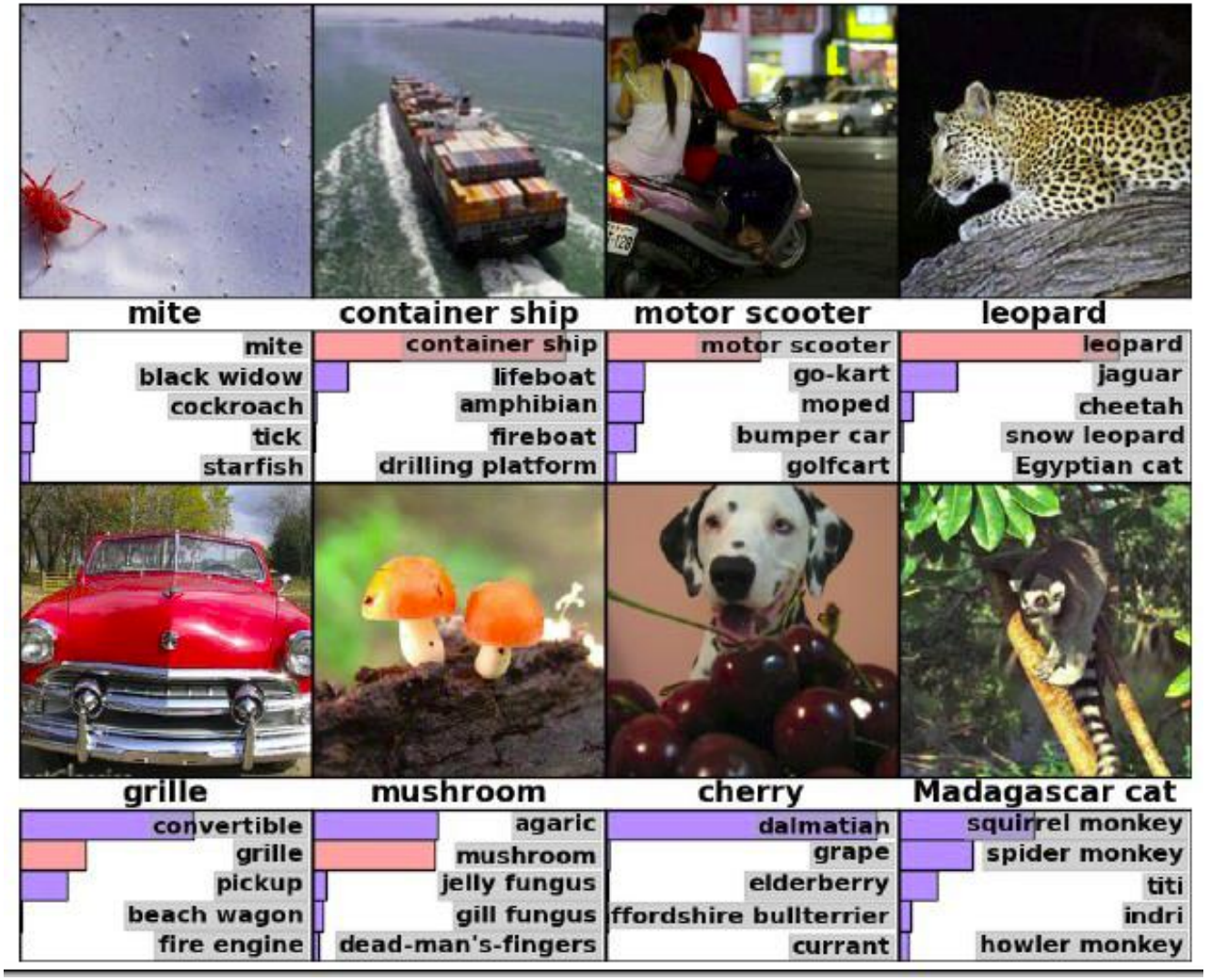
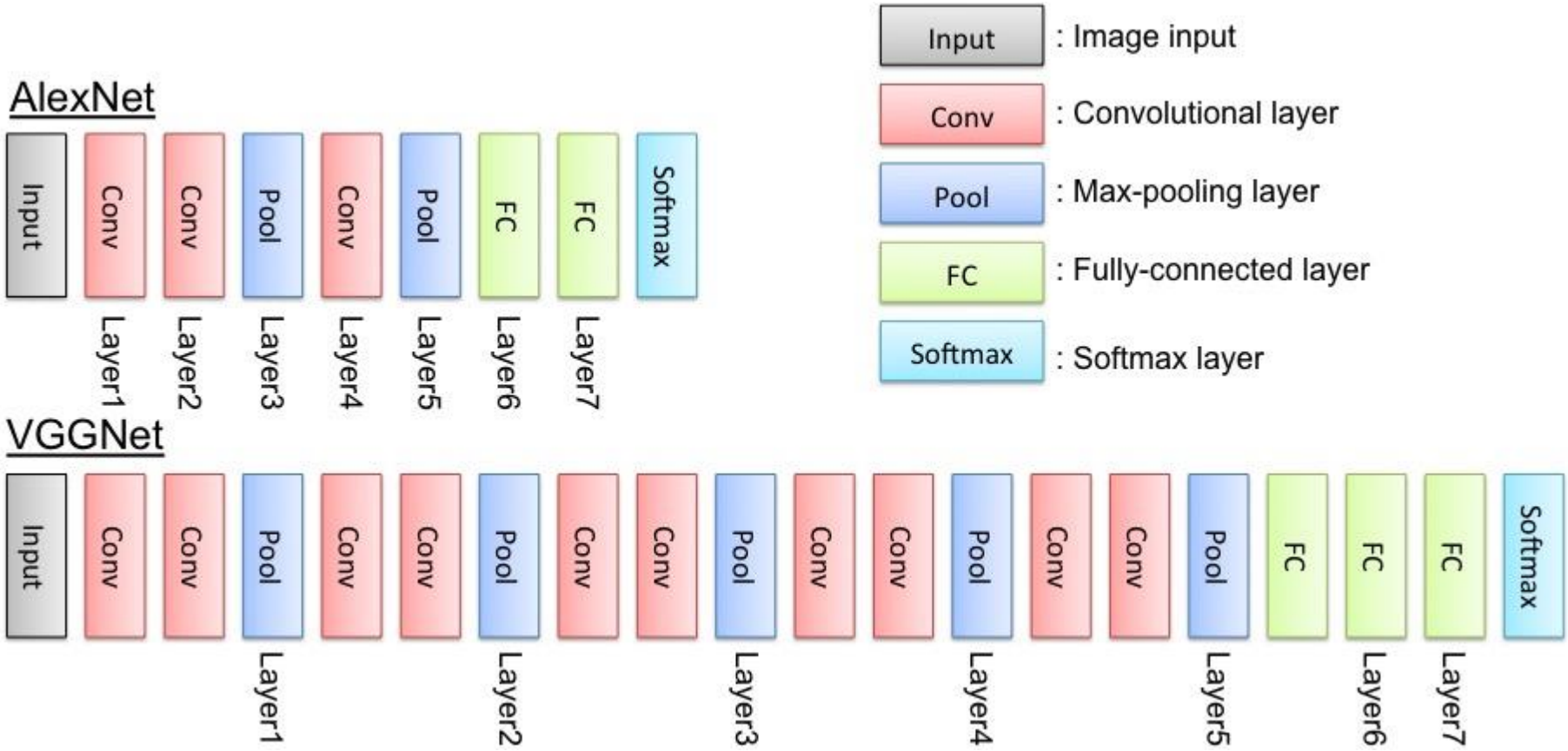


Image source: A. Krizhevsky, I. Sutskever and G.E. Hinton, NIPS 2012

CNN Architectures: VGGNet (2014/15)



K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, ICLR 2015

Image source: Hirokatsu Kataoka

CNN Architectures: VGGNet (2014/15)

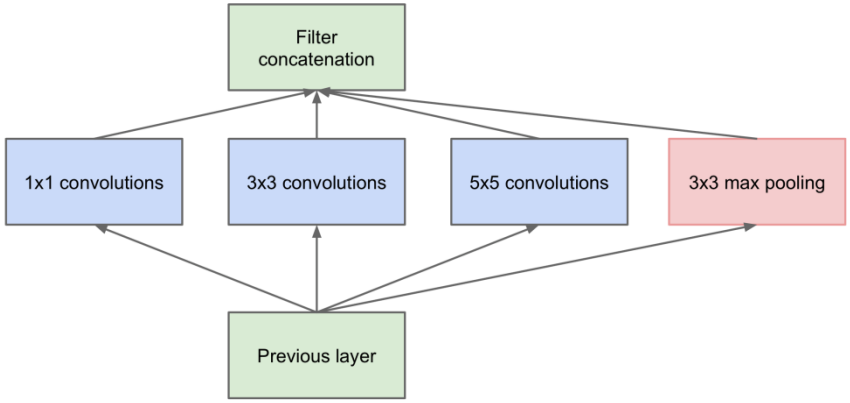
- **Main ideas**
  - Deeper network
  - Stacked convolutional layers with smaller filters (+ nonlinearity)
  - ReLU nonlinearities
  - Detailed evaluation of all components
- **Results**
  - Improved ILSVRC top-5 error rate to 6.7%.

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	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 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

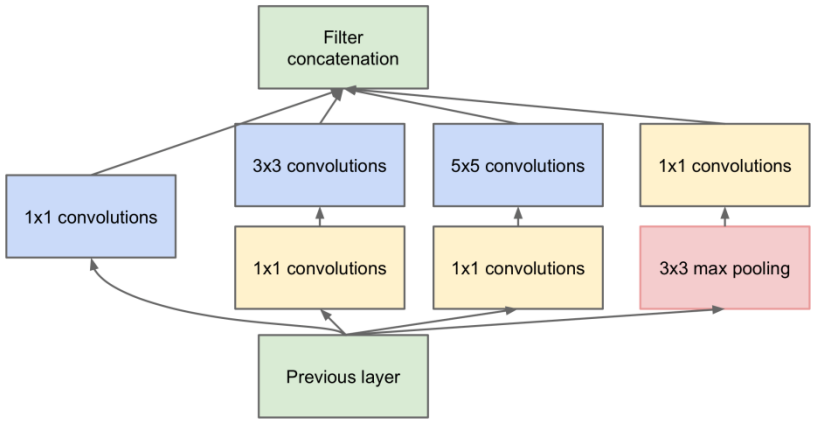
Mainly used

Image source: Simonyan & Zisserman

CNN Architectures: GoogLeNet (2014)



(a) Inception module, naïve version

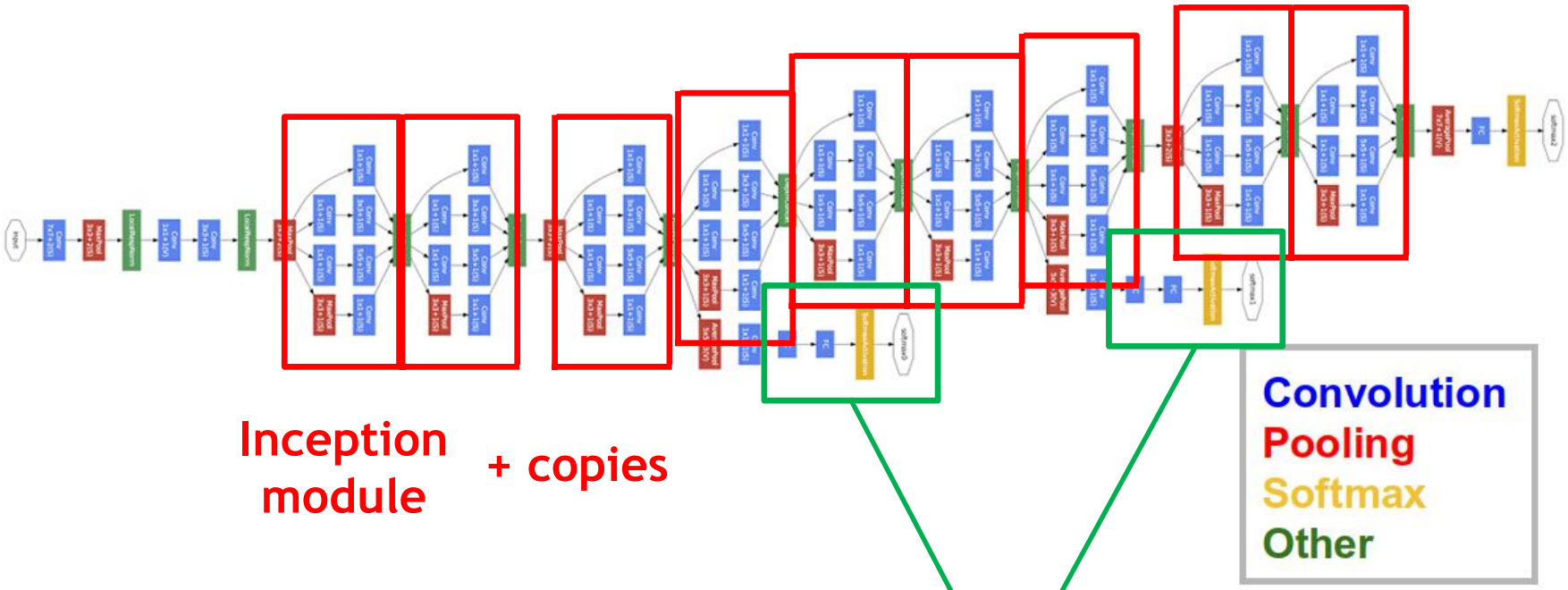


(b) Inception module with dimension reductions

- **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.

GoogLeNet Visualization



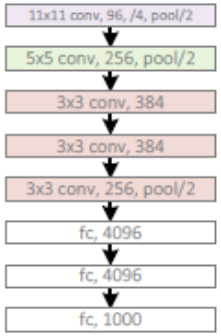
Results on ILSVRC

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)	-	-	8.1
MSRA (He et al., 2014) (1 net)	27.9	9.1	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	-

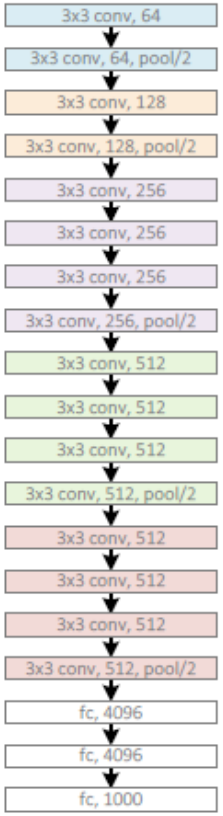
Image source: Simonyan & Zisserman

Newest Development: Residual Networks

AlexNet, 8 layers
(ILSVRC 2012)



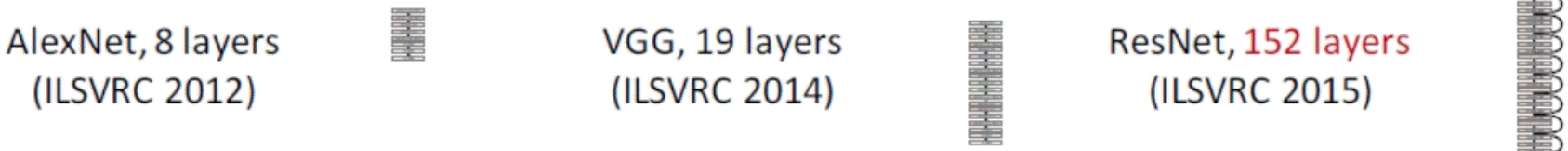
VGG, 19 layers
(ILSVRC 2014)



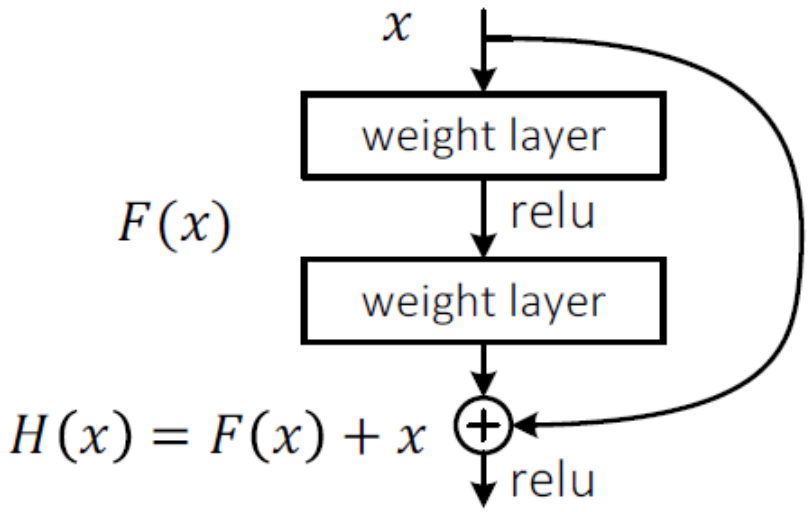
GoogleNet, 22 layers
(ILSVRC 2014)

Slide credit: Kaiming He

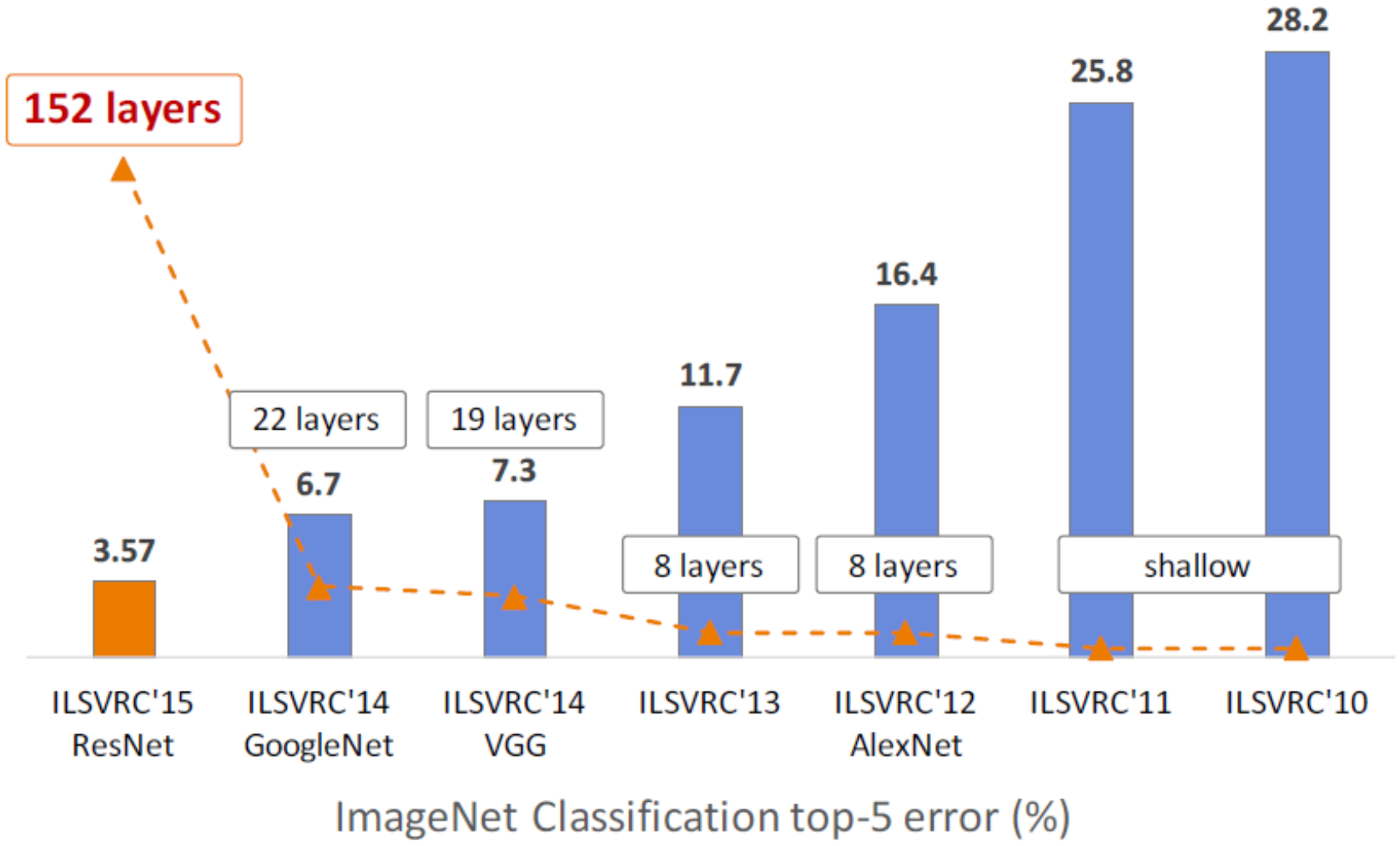
Newest Development: Residual Networks



- **Core component**
  - **Skip connections bypassing each layer**
  - **Effect: better propagation of gradients to the deeper layers**
  - **This makes it possible to train very deep networks.**



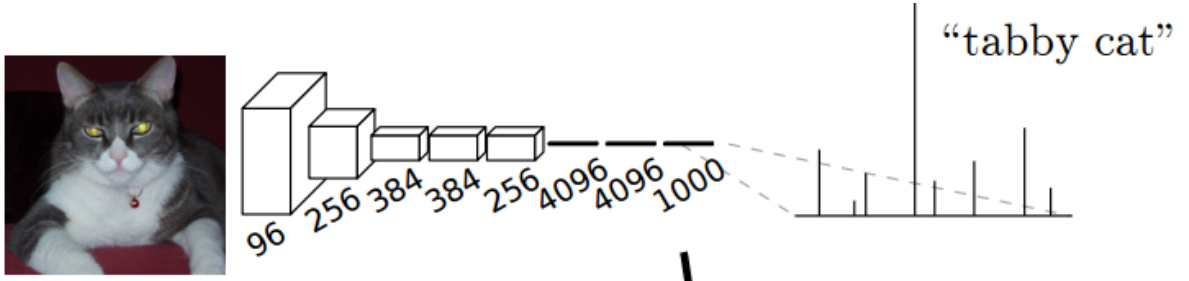
ImageNet Performance



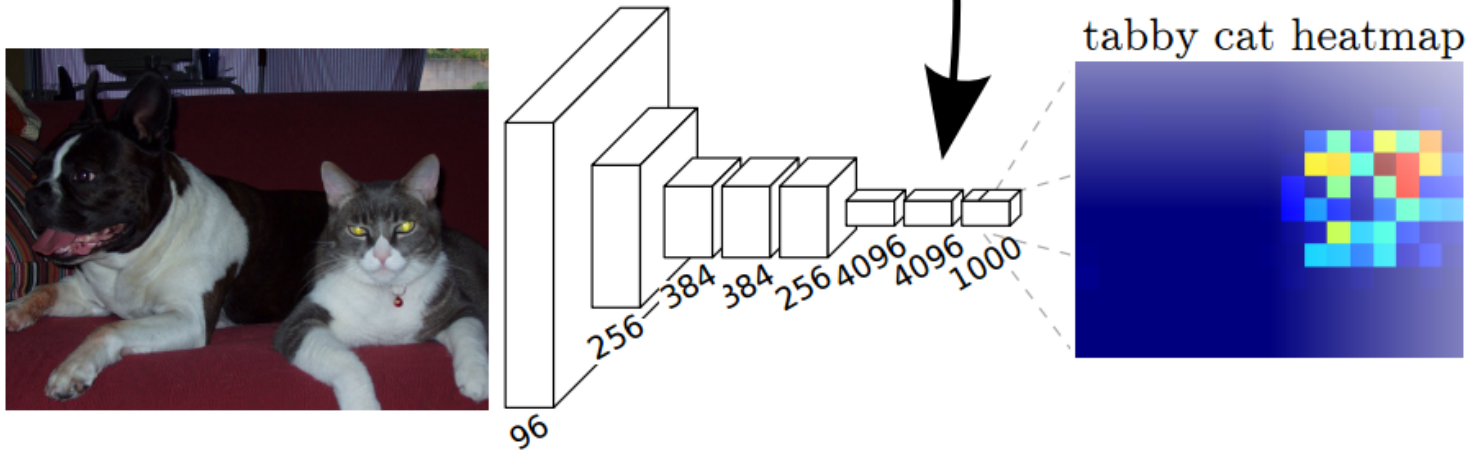
Slide credit: Kaiming He

Fully Convolutional Networks

- CNN



- FCN



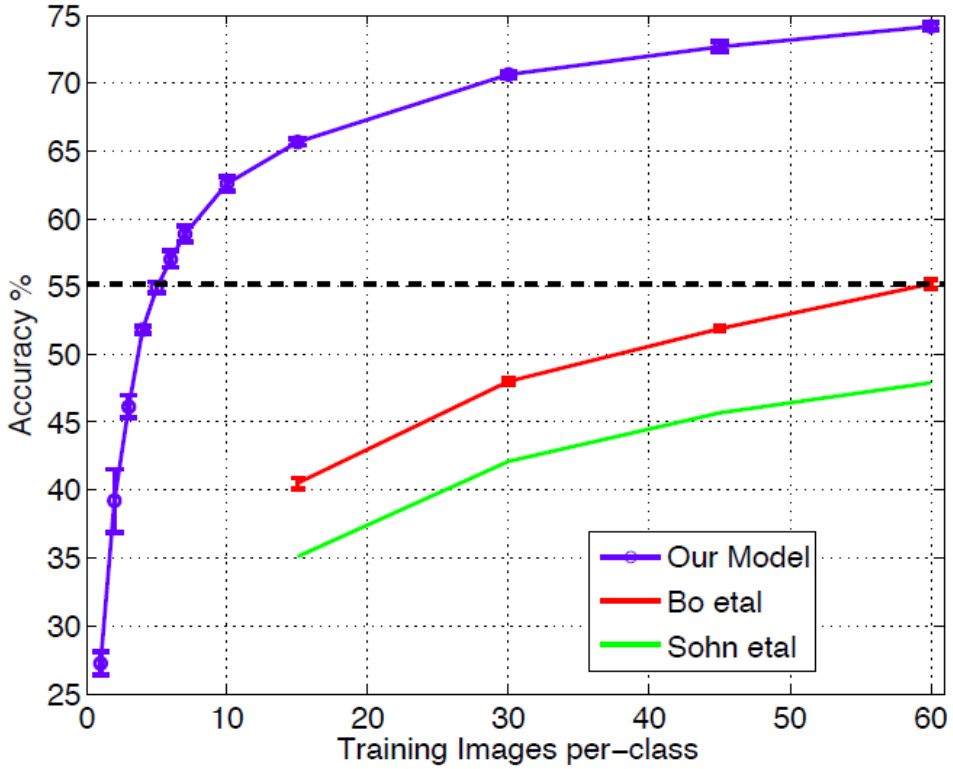
- Intuition
  - Think of FCNs as performing a sliding-window classification, producing a heatmap of output scores for each class
  - This makes it possible to process images of arbitrary size.

Image source: Long, Shelhamer, Darrell

Outline

- Recap: CNNs
  - Convolutional layers
  - Pooling layers
- CNN Architectures and Historical Development
  - LeNet
  - AlexNet
  - VGGNet
  - GoogLeNet
  - ResNet
  - Fully Convolutional Networks
- **Applications**
  - **Object Detection**
  - **Semantic Image Segmentation**
  - **Matching**

The Learned Features are Generic



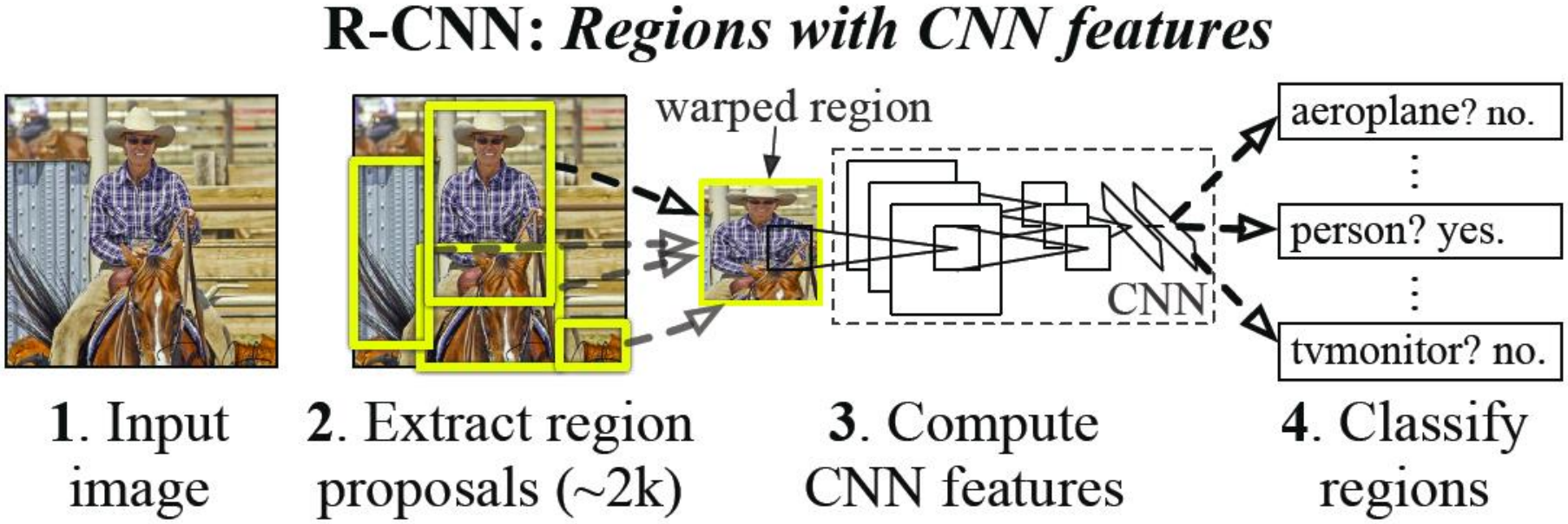
state of the art level (pre-CNN)

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

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

Image source: M. Zeiler, R. Fergus

Other Tasks: Object Detection



- **Results on PASCAL VOC 2010 Detection benchmark**
  - **Pre-CNN state of the art: 33.4% mAP DPM**
    40.4% mAP SegDPM
  - **R-CNN: 53.7% mAP**

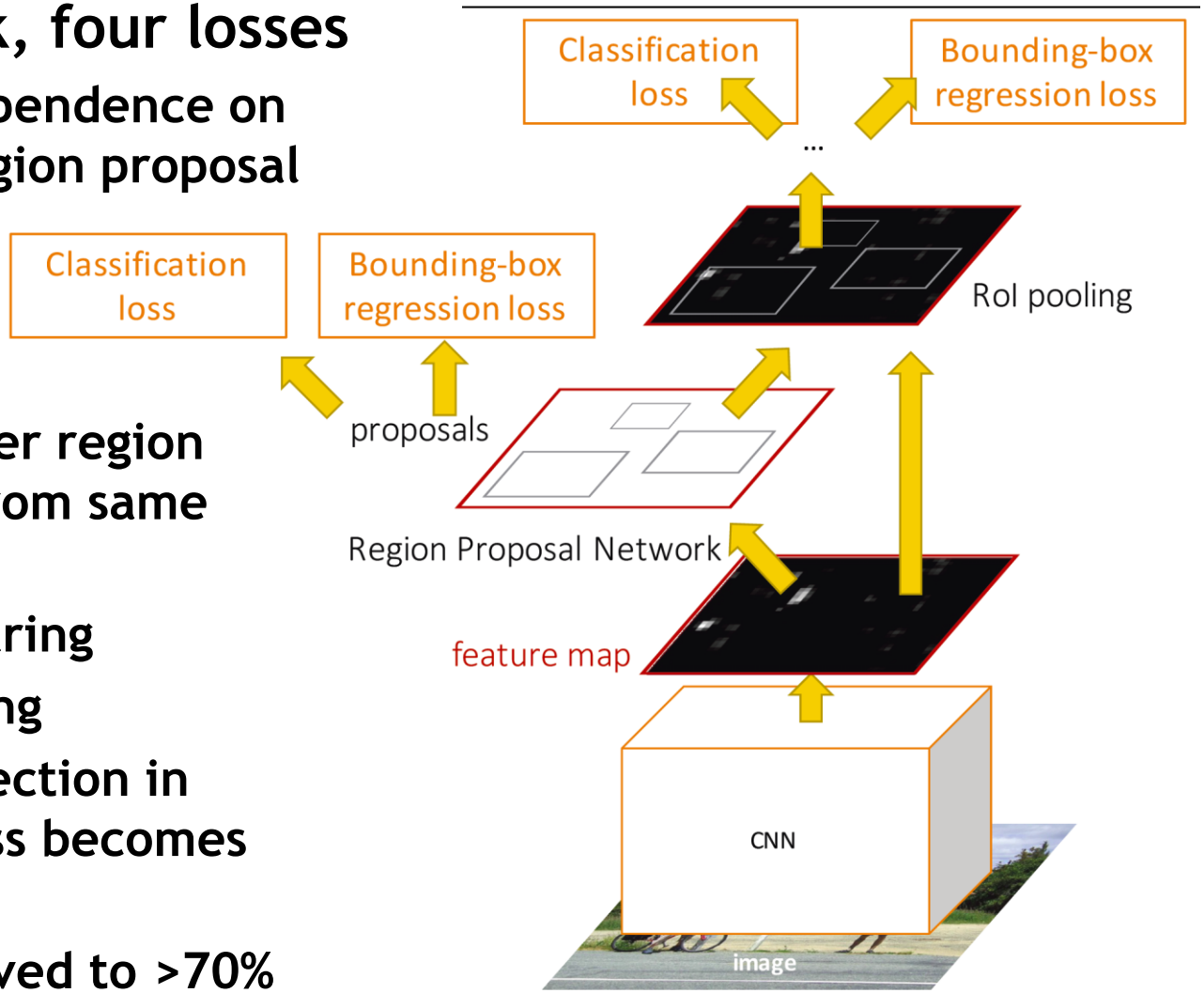
R. Girshick, J. Donahue, T. Darrell, and J. Malik, Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, CVPR 2014

Most 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

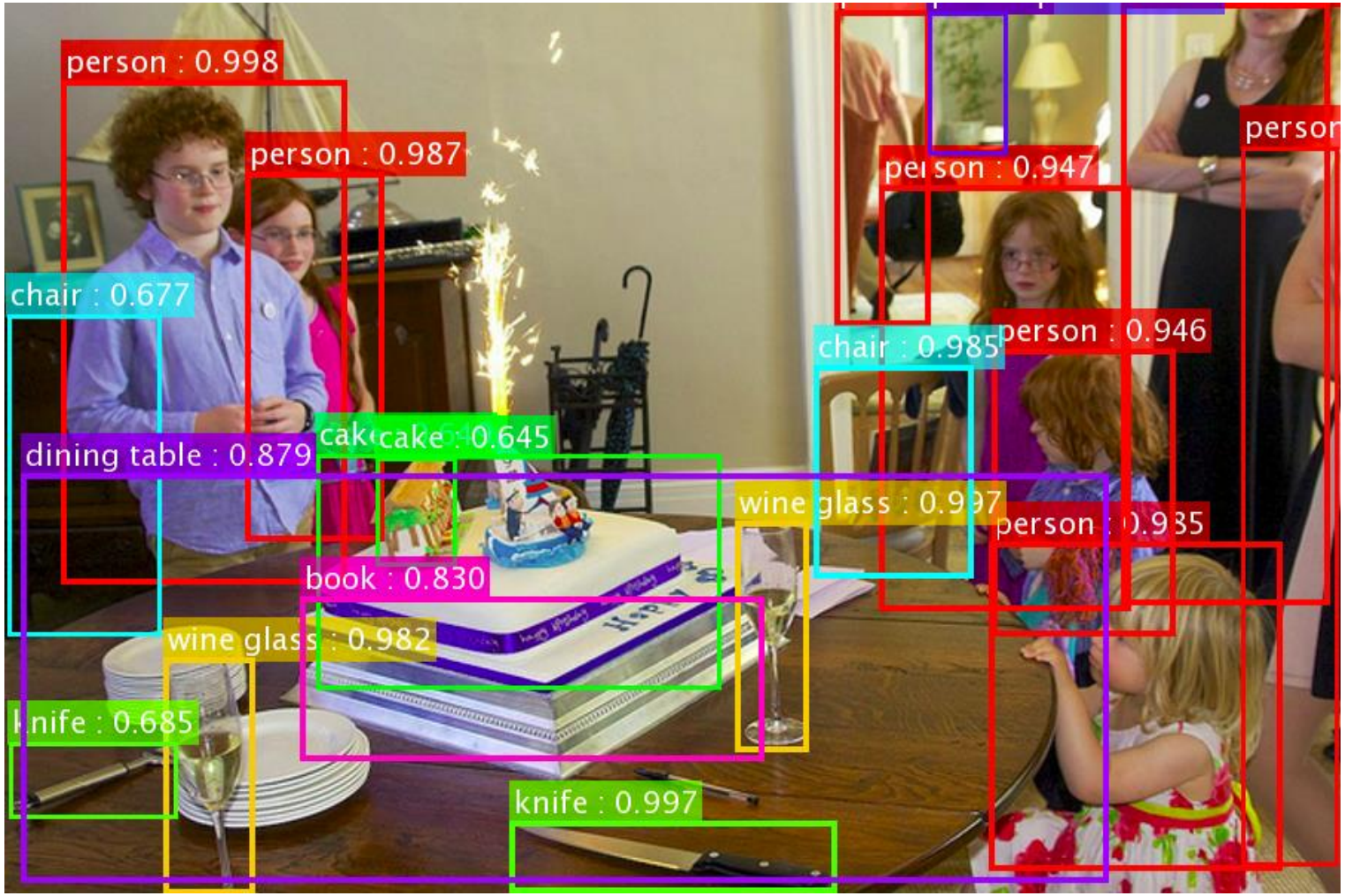
⇒ Object detection in a single pass becomes possible.

⇒ mAP improved to >70%



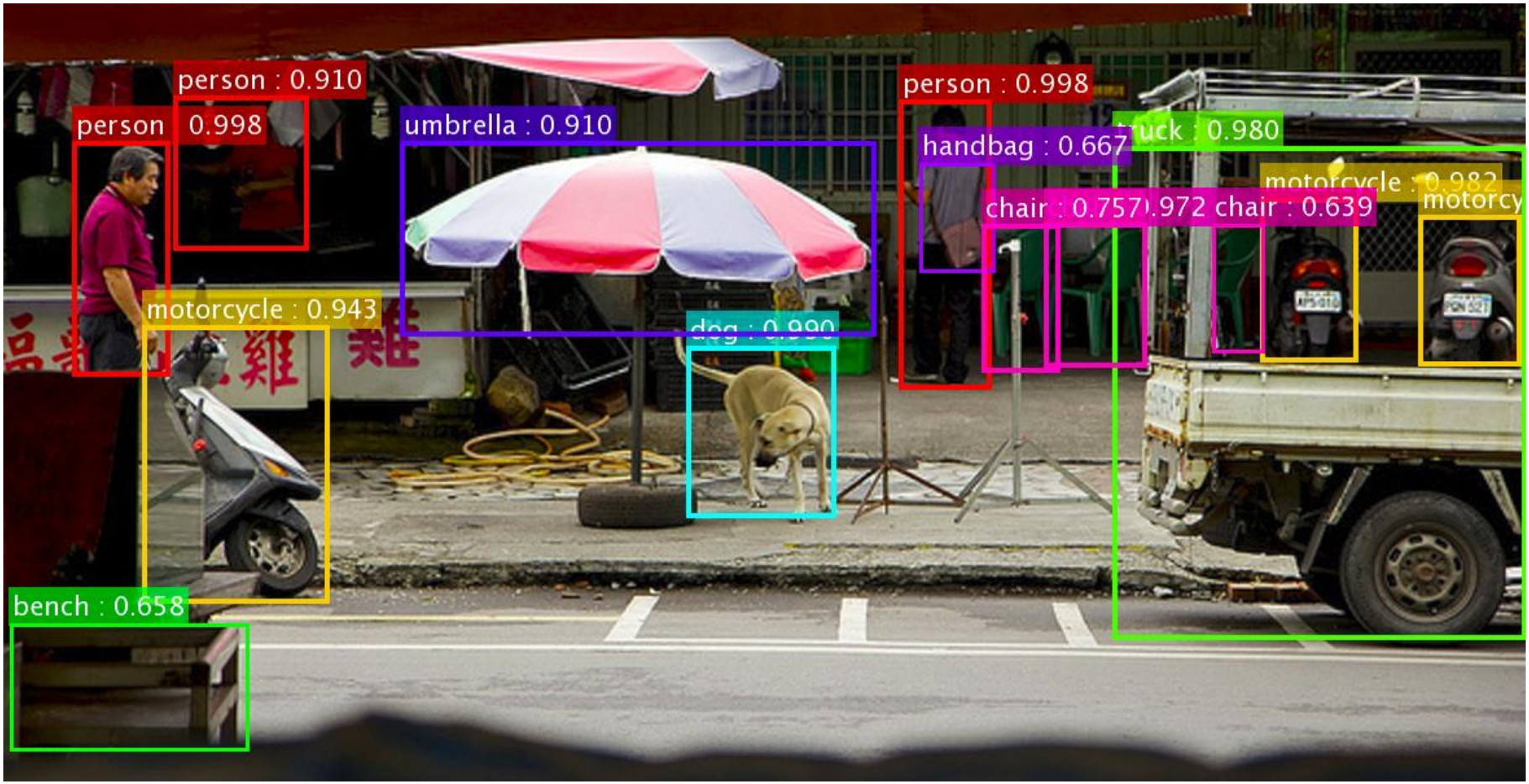
Slide credit: Ross Girshick

Faster R-CNN Results (using ResNets)



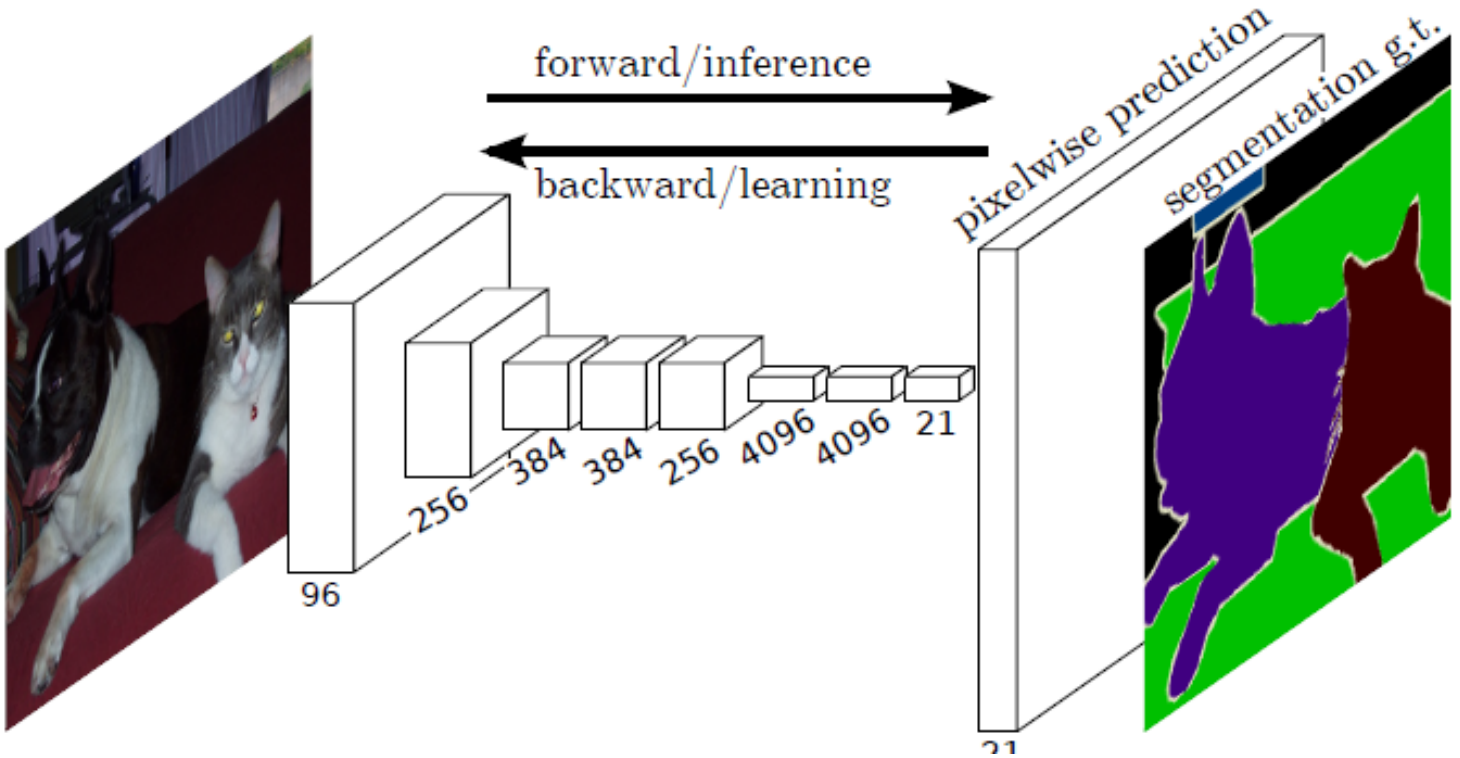
K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, CVPR 2016.

Faster R-CNN Results (using ResNets)



K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, CVPR 2016.

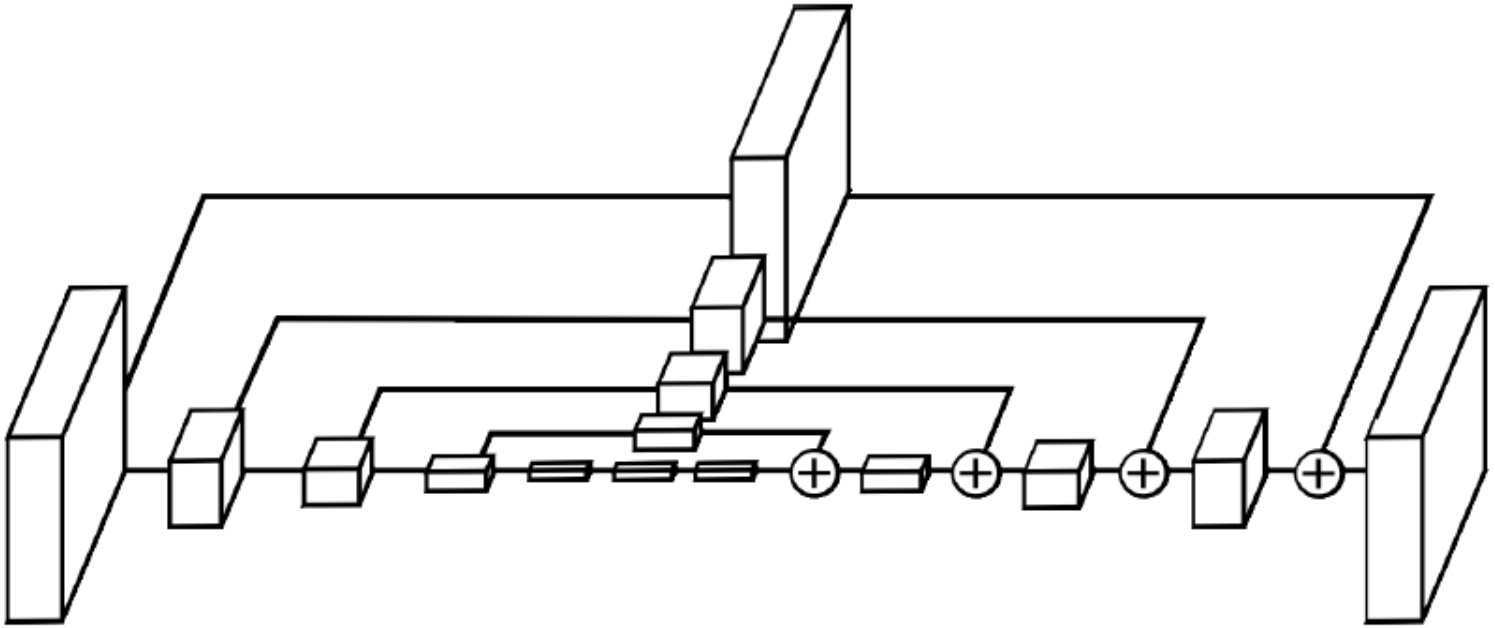
Semantic Image Segmentation



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

Image source: Long, Shelhamer, Darrell

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

Image source: Newell et al.

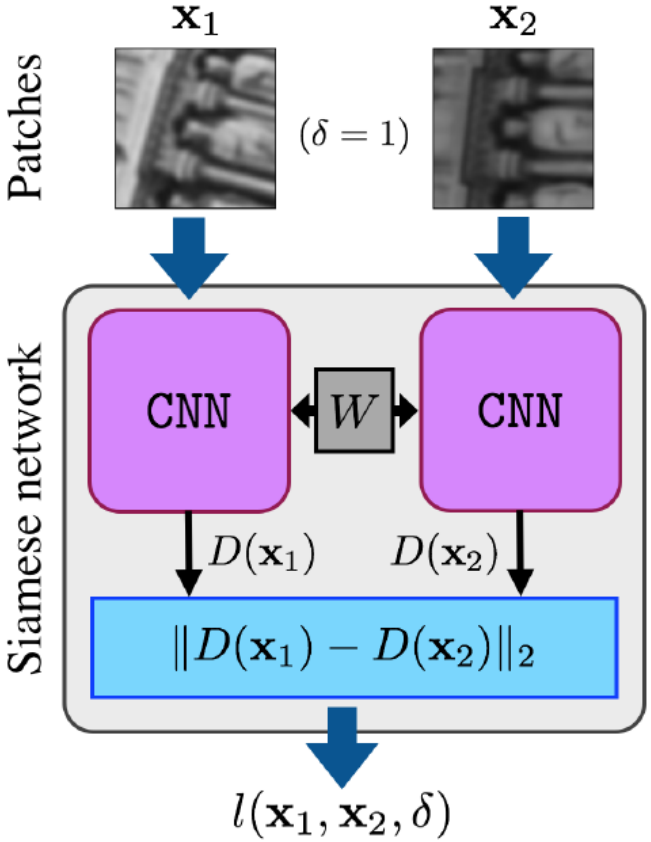
Semantic Segmentation Results

- State-of-the-art results on Cityscapes benchmark
  - Own work, based on an extension of ResNets

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

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
  - …



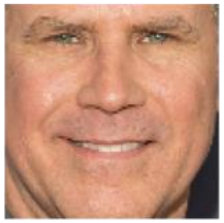
Extension: Triplet Loss Networks

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

Negative

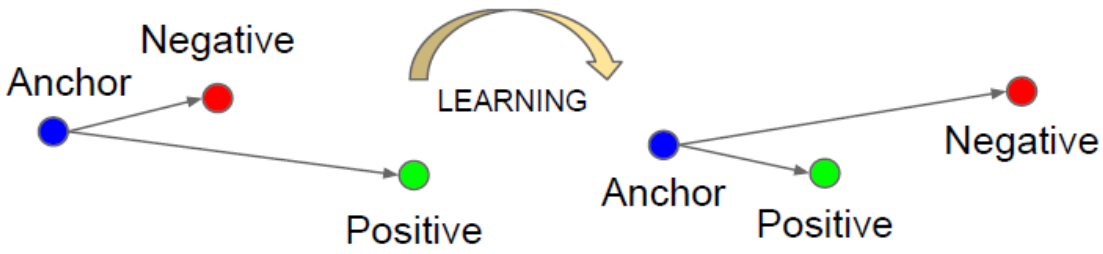
Anchor

Positive



  - Apply triplet loss to learn an embedding $f(\cdot)$ that groups the positive example closer to the anchor than the negative one.

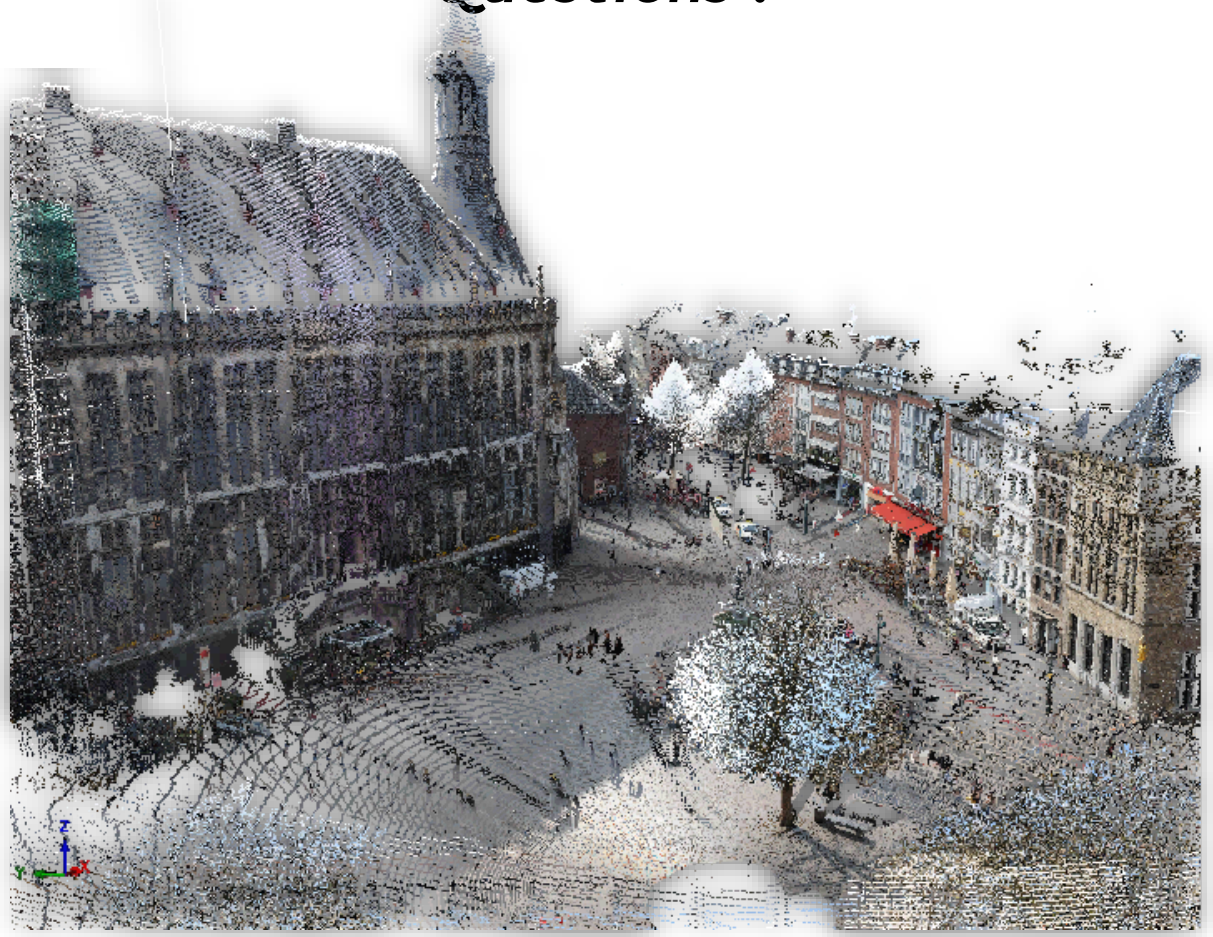
$$\|f(x_i^a) - f(x_i^p)\|_2^2 < \|f(x_i^a) - f(x_i^n)\|_2^2$$



⇒ Used with great success in Google's FaceNet face recognition

Thank you very much!

Questions ?



http://www.vision.rwth-aachen.de/

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.

References and Further Reading

- **ResNet**
  - K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, CVPR 2016.