Boston University
CAS CS 640: AI

Lecture on

Introduction to Computer Vision

Mahir Patel and Margrit Betke

October 17, 2024

Learning Objectives for this Lecture

Computer Science

- Understand formats of images used as inputs to AI models: greyscale, color, medical scans
- Understand differences and similarities between pre-2012 “traditional computer vision” and post-2012 neural-network-based computer vision & see examples
- Understand why convolution is powerful
- Understand the connection between convolution and correlation
- Understand template matching with image pyramids
- Understand CNNs as a learning hierarchy of features
- Learn about early CNN used in computer vision: LeCun’s work on recognizing handwritten numbers
- Understand CNN concepts, e.g., convolution layers, fully connected (dense) layers, non-linearity (ReLU), pooling (downsampling)

Multi-Resolution Matching

BOSTON UNIVERSITY
Computer Science

Normalized correlation coefficient over multi-resolution search space:

$$r = \frac{1}{n} \frac{\Sigma_i (s_i - mean(s))\,(m_i - mean(m))}{(\sigma_s\,\sigma_m)}$$

←Template matched over all resolutions →

Finding the Face and its Movement by Locating the Best Match of a Face Template

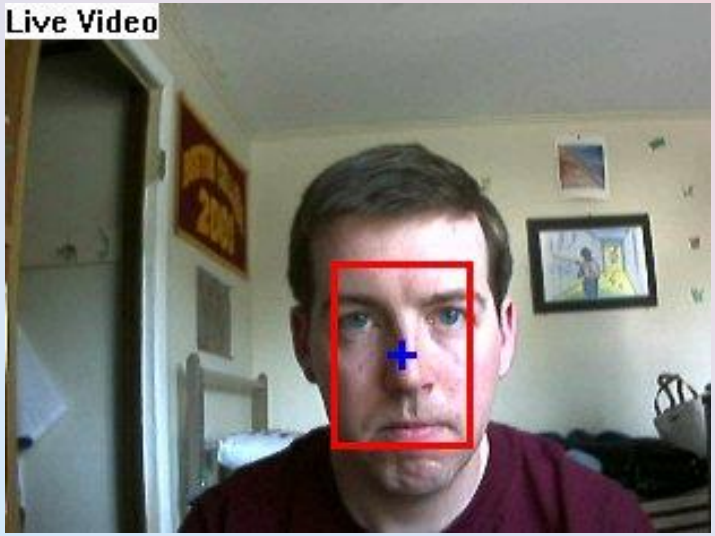
(a) Input

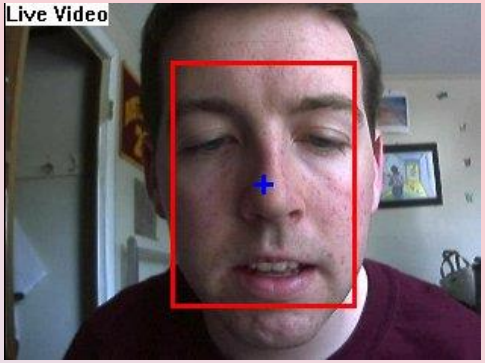
You can apply template matching to a small version of your input image and use that search result to start searching for a match in the 2nd smallest images. Repeat until the original size is processed.

(d) Correlation

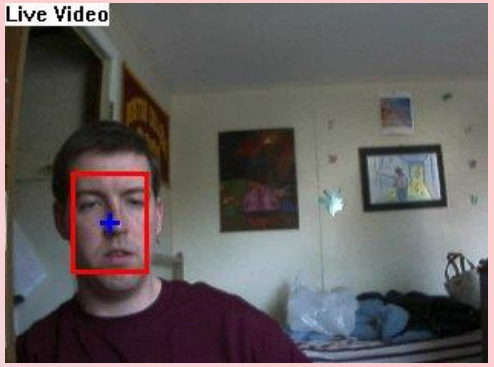
Face Detection

Data Variability



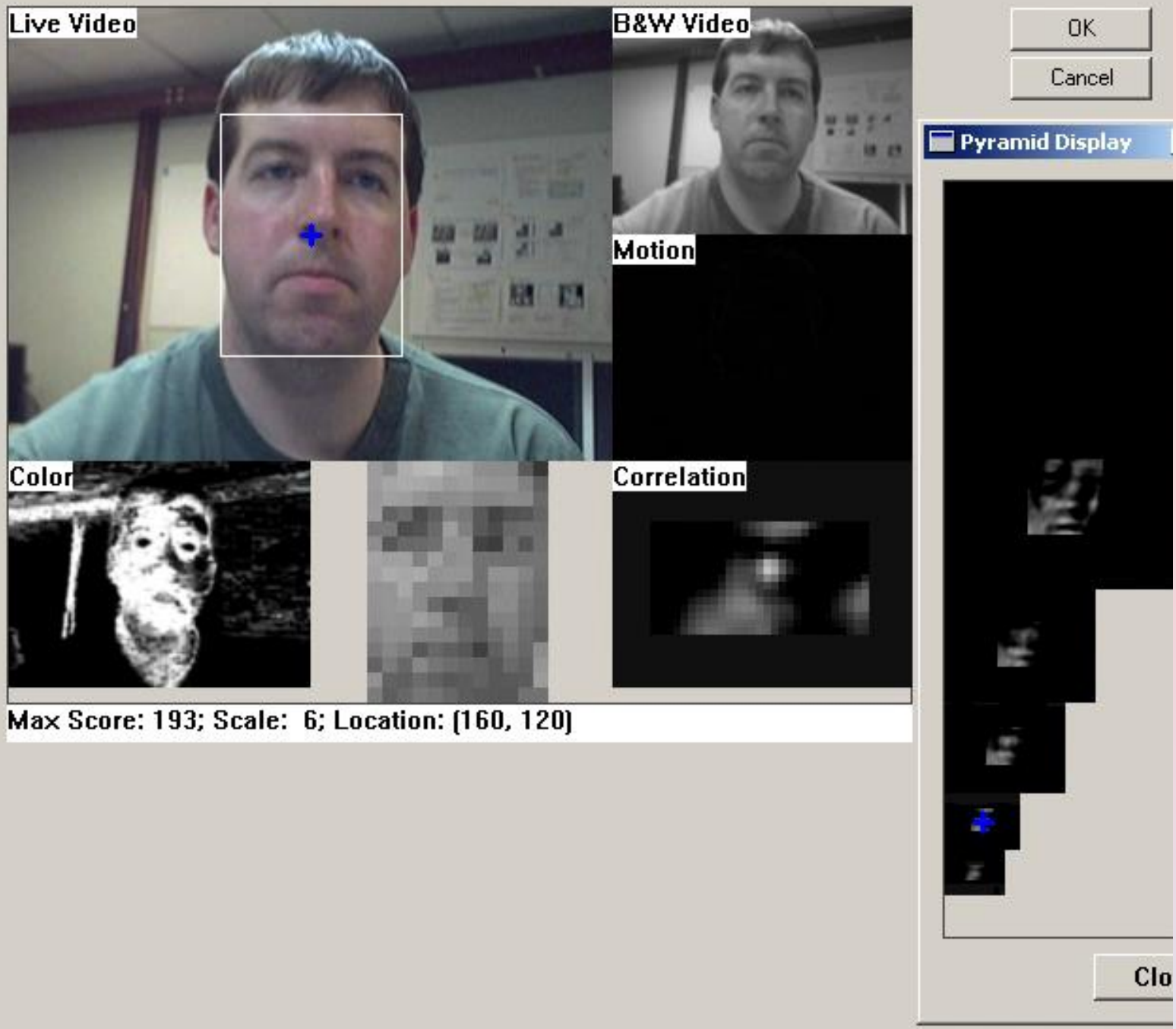


Large Face



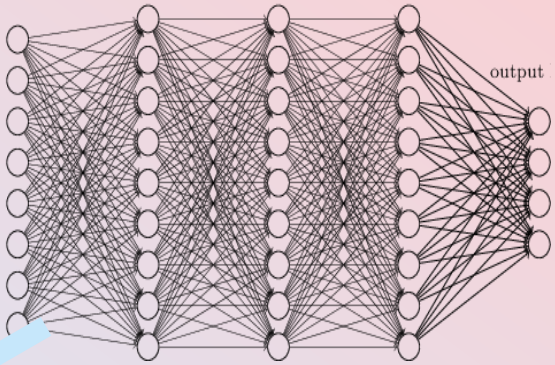
Small Face

Shadows
Cluttered background


Live Video
B&W Video
OK
Cancel
Pyramid Display
Motion
Color
Correlation
Max Score: 193; Scale: 6; Location: (160, 120)
Clo

Back to Neural Nets & their Success in Solving Computer Vision Problems

Large labeled datasets



Deep neural networks

GPU technology

Slide credit: Dinesh Jayaraman

Convolutional Neural Networks (CNN, ConvNet, DCN)

- CNN = a multi-layer neural network with
  - **Local** connectivity:
    - Neurons in a layer are only connected to a small region of the layer before it
  - **Share** weight parameters across spatial positions:
    - Learning shift-invariant filter kernels

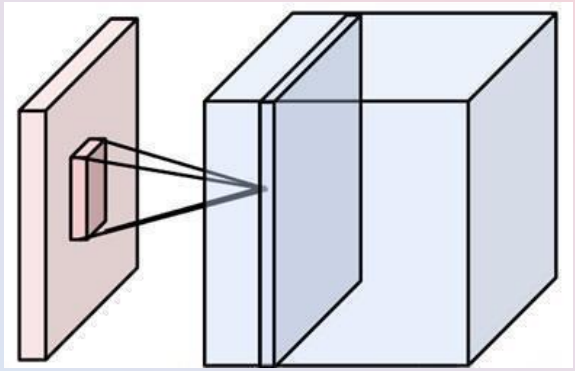


Image credit: A. Karpathy

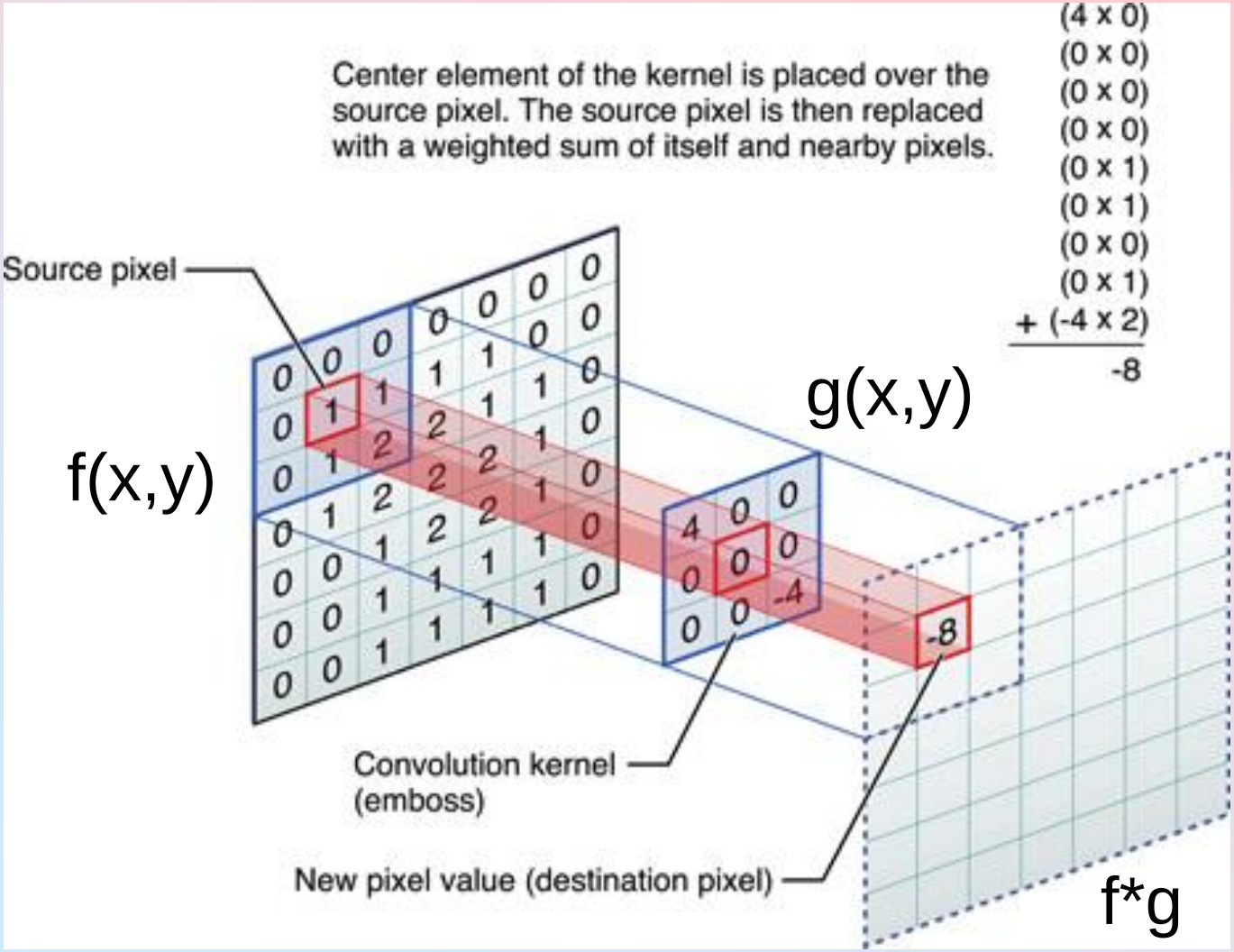
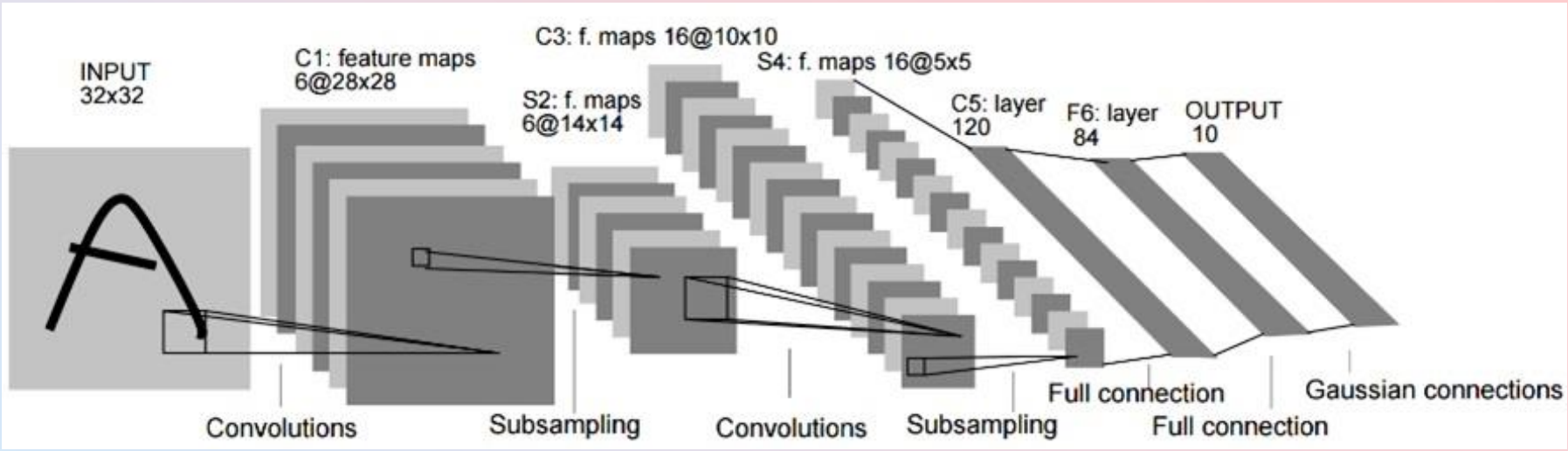


Image Credit: Madhushree Basavarajaiah

LeNet [LeCun et al.]



1990: Zipcode recognition

http://yann.lecun.com/exdb/lenet/multiples.html

LeNet-1 from 1993

Gradient-based learning applied to document recognition [LeCun, Bottou, Bengio, Haffner 1998]

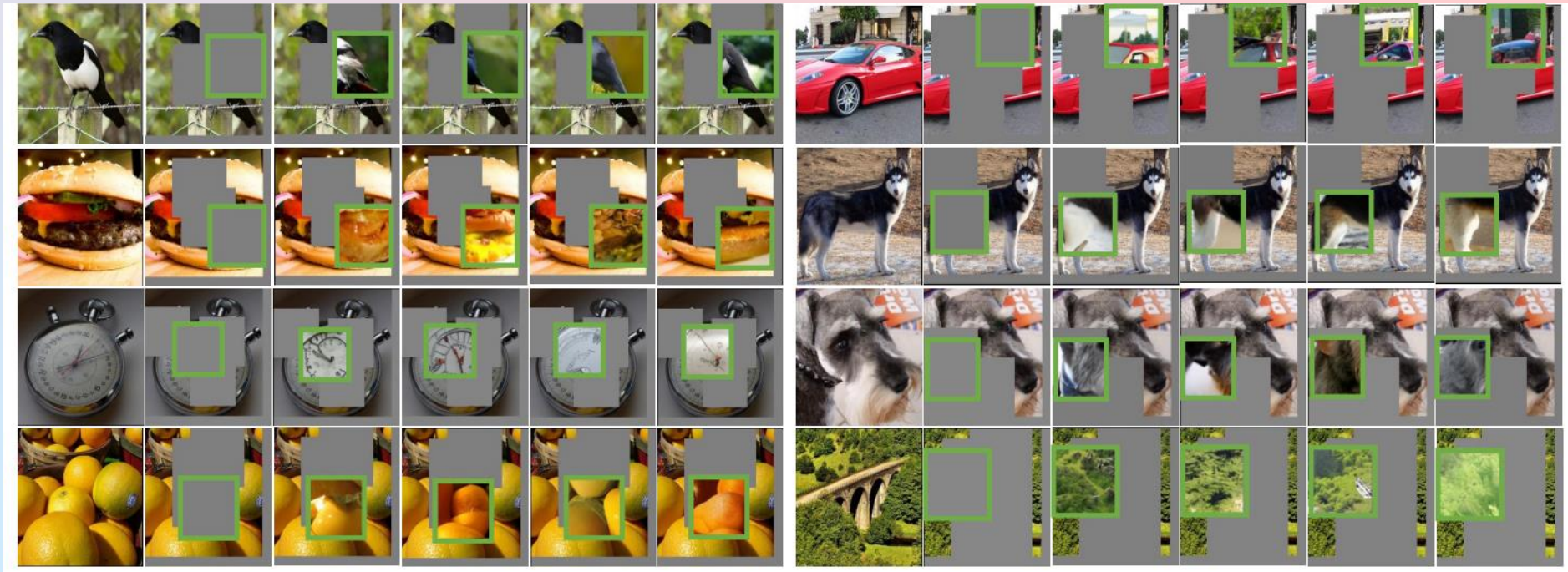
LeCun Interview, Oct. 5, 2023

- https://www.rsipvision.com/ICCV2023-Thursday/

Yann LeCun

- VP and Chief AI Scientist, Facebook
- Silver Professor of Computer Science, Data Science, Neural Science, and Electrical and Computer Engineering, New York University
- ACM Turing Award Laureate
- Member, National Academy of Engineering

LeCun's 2023 Focus: Predict Content of Masked-out Images/Video Frames



I-JEPA

Image Credit: 2301.08243.pdf (arxiv.org)

LeCun's focus: Predict Content of Masked-out Images/Video Frames

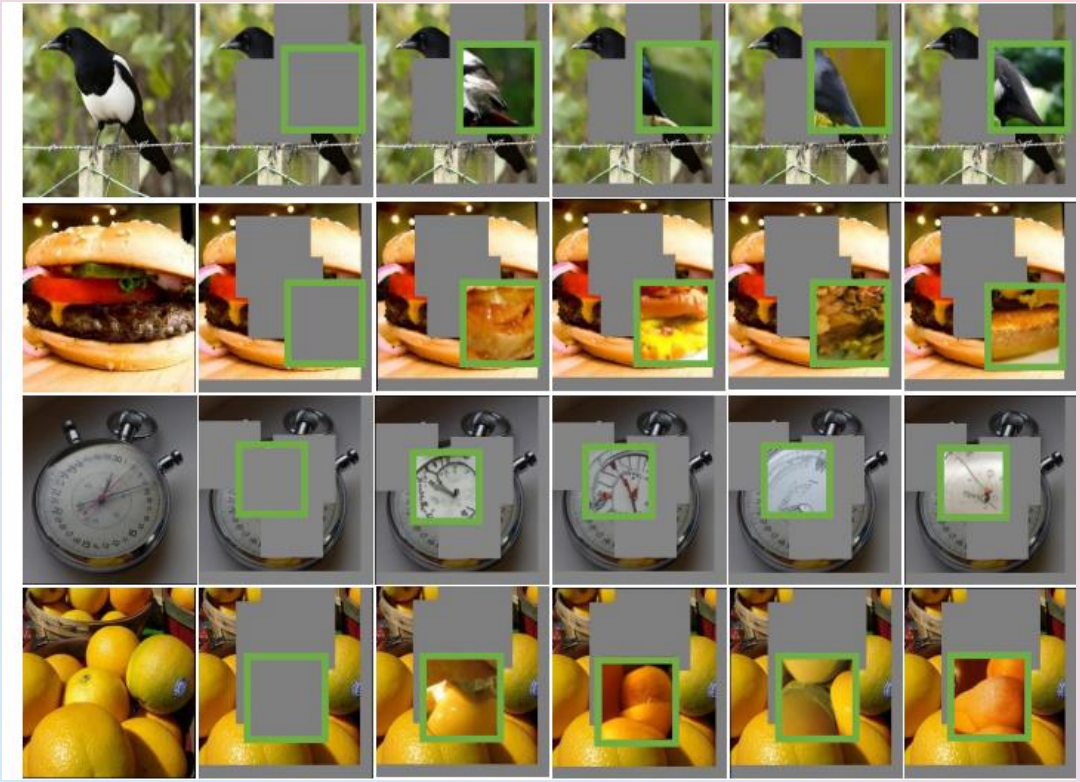
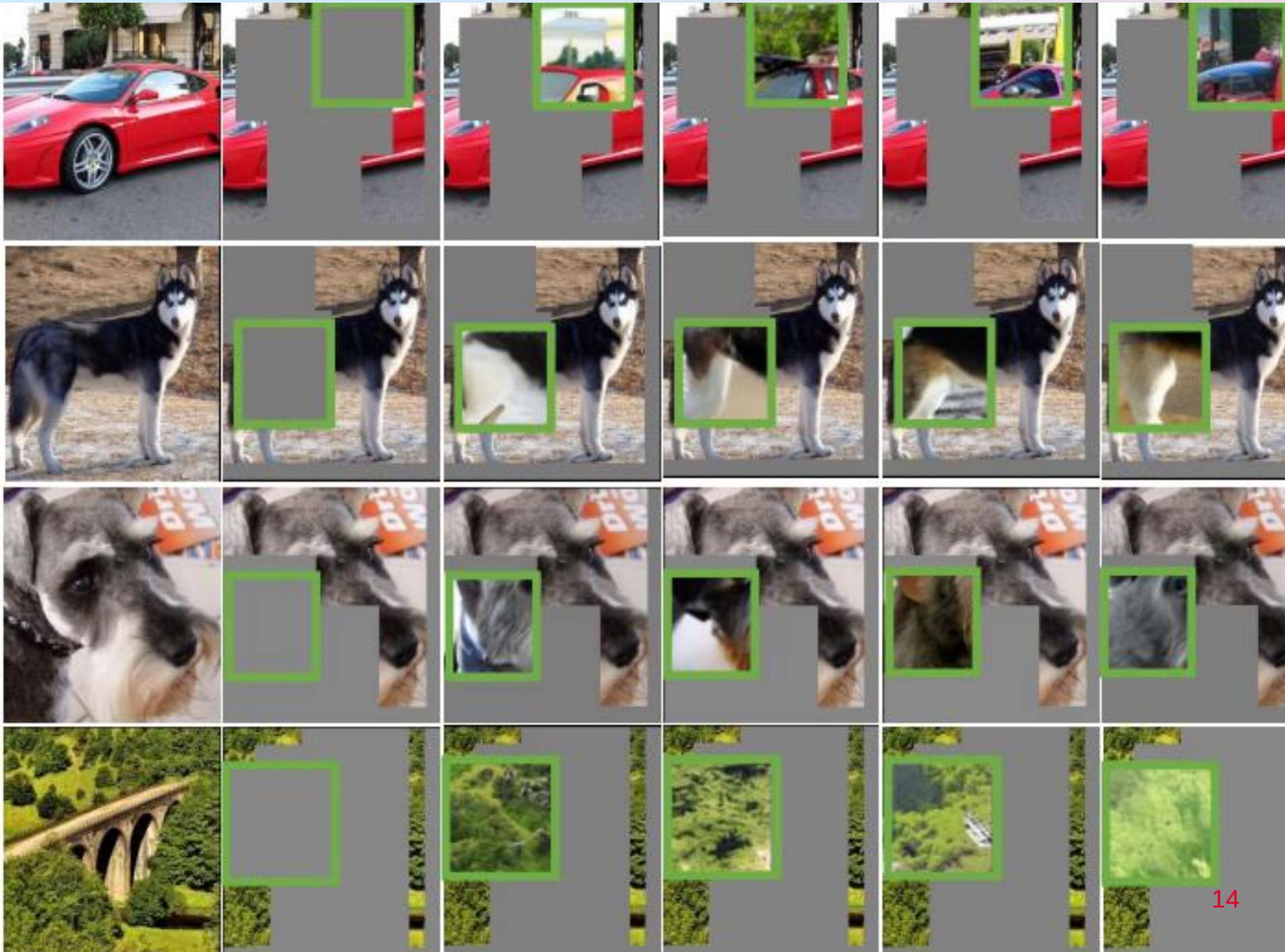
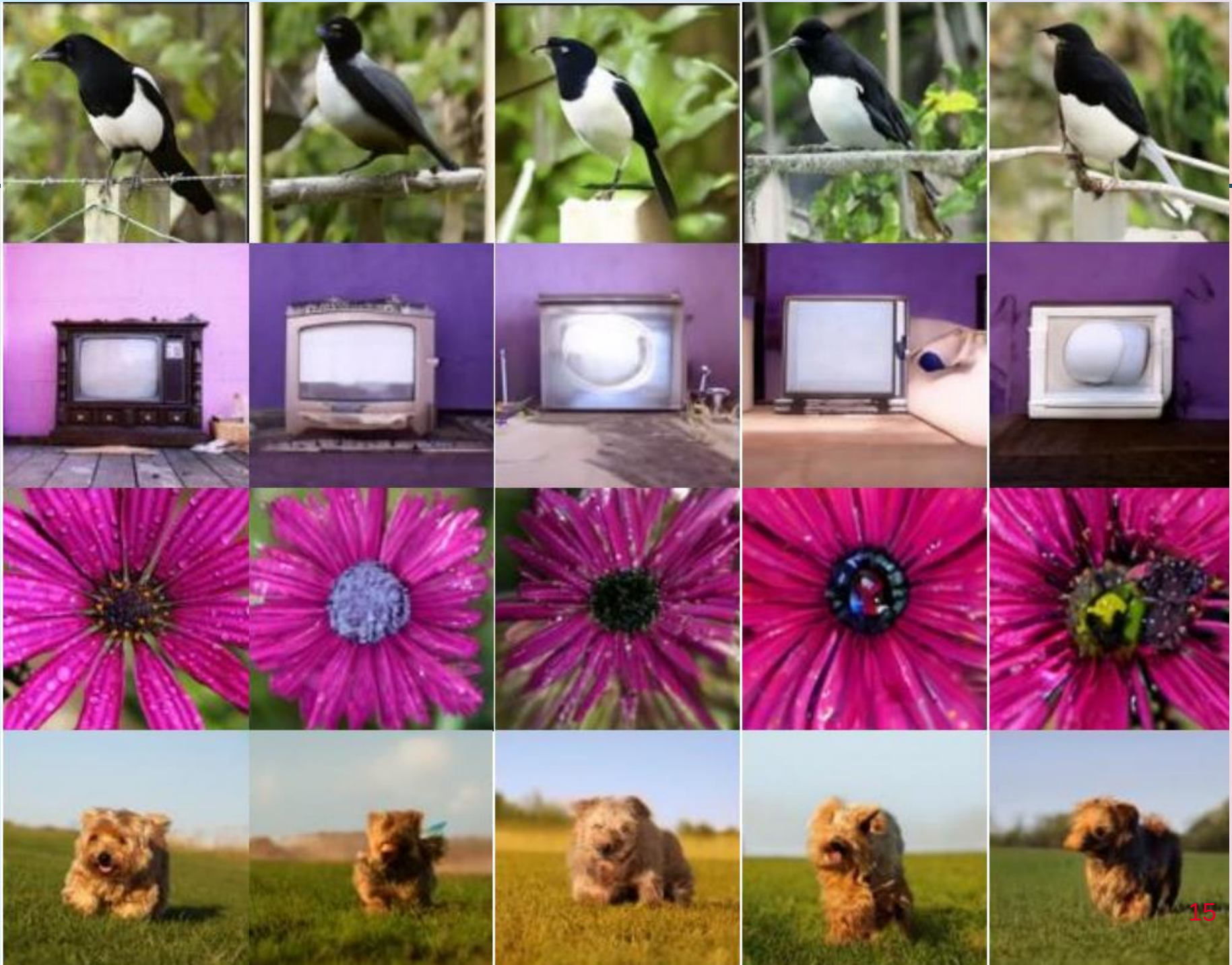


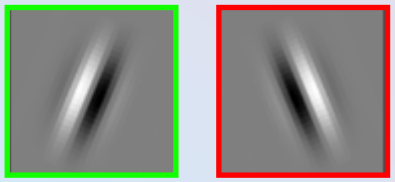
Image Credit: 2301.08243.pdf (arxiv.org)





Another example of 2D Convolution

- Weighted moving sum



What do you think each filter does?

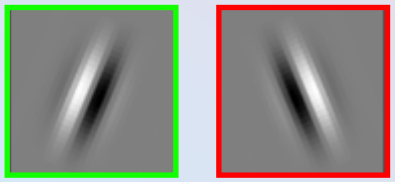
Input

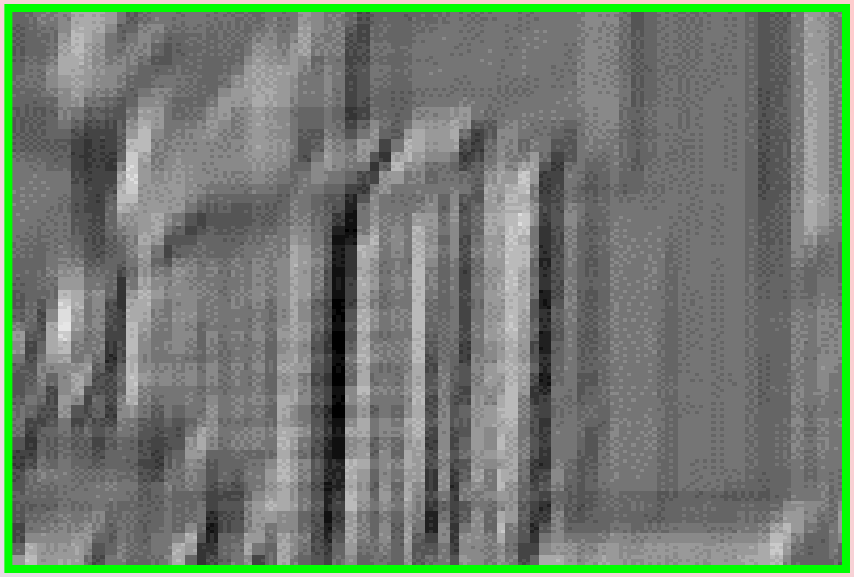
Feature Activation Map

slide credit: S. Lazebnik

Another example of 2D Convolution

- Weighted moving sum



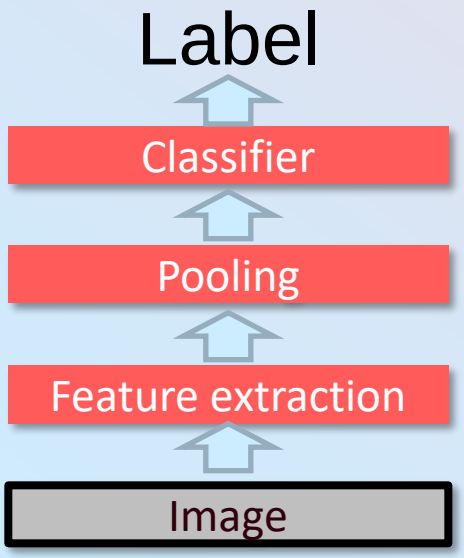


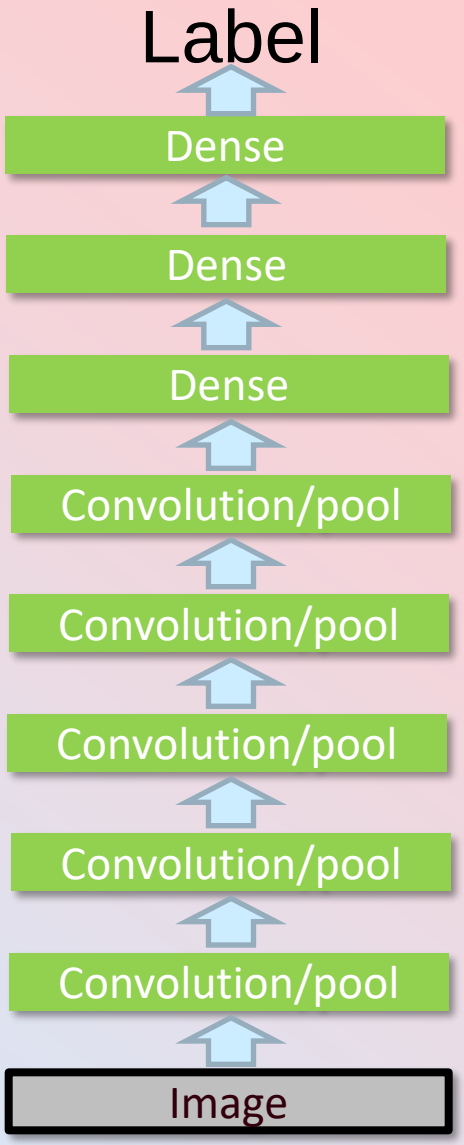
Input

Feature Activation Map

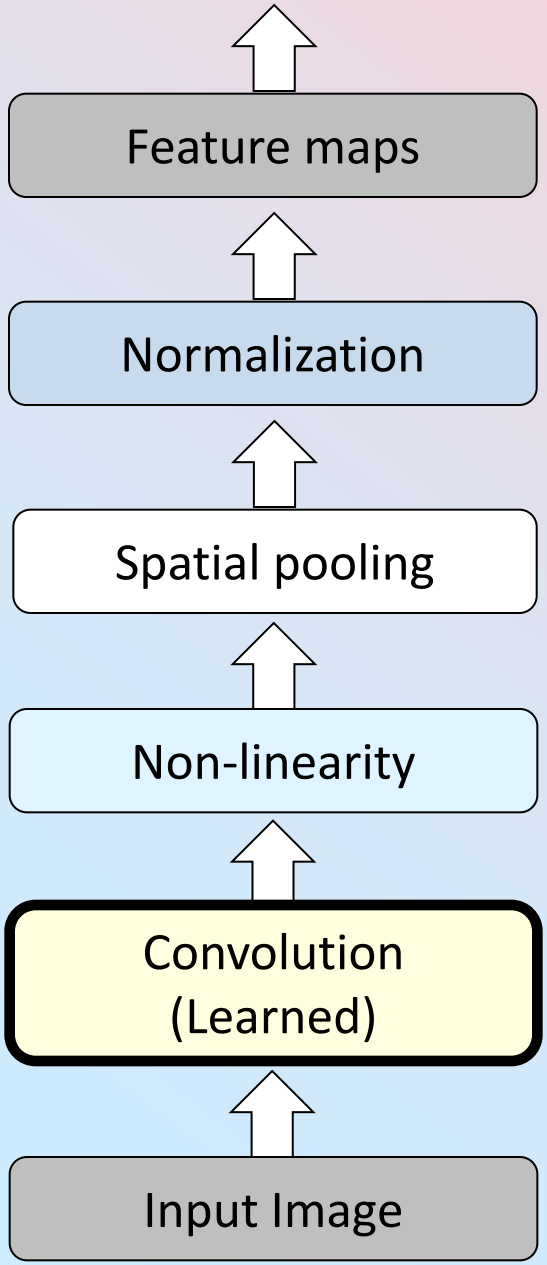
slide credit: S. Lazebnik

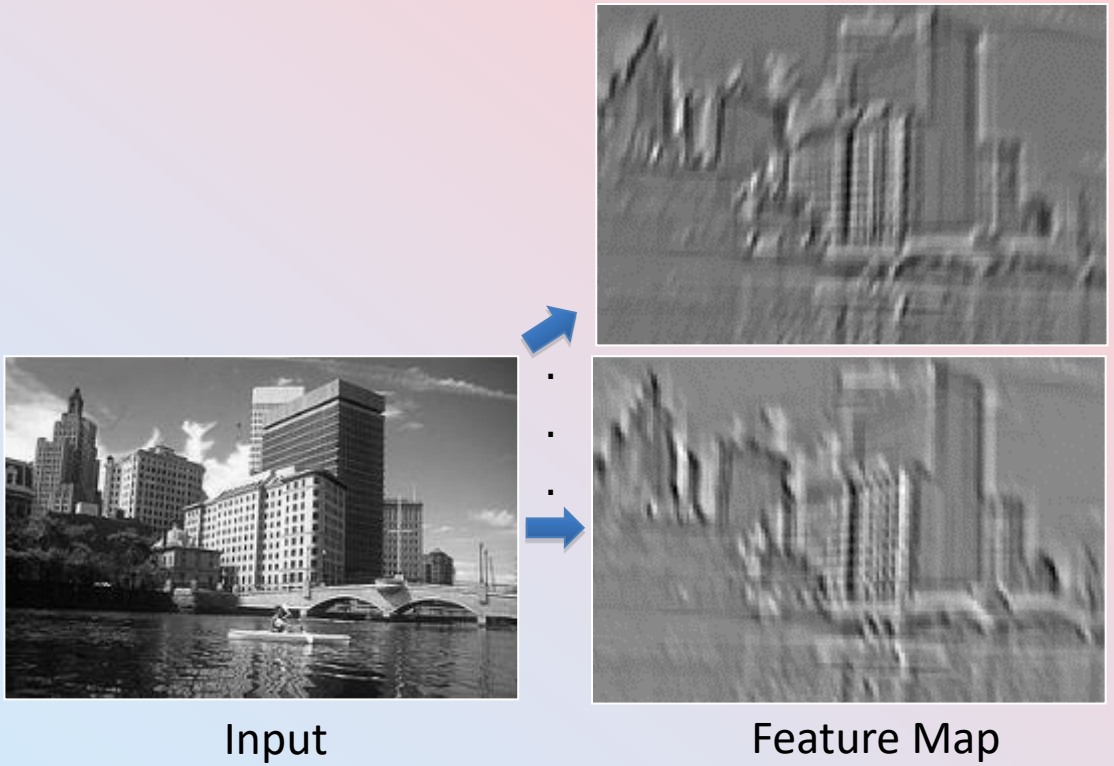
Traditional versus NN-based Computer Vision: Engineered versus Learned Features





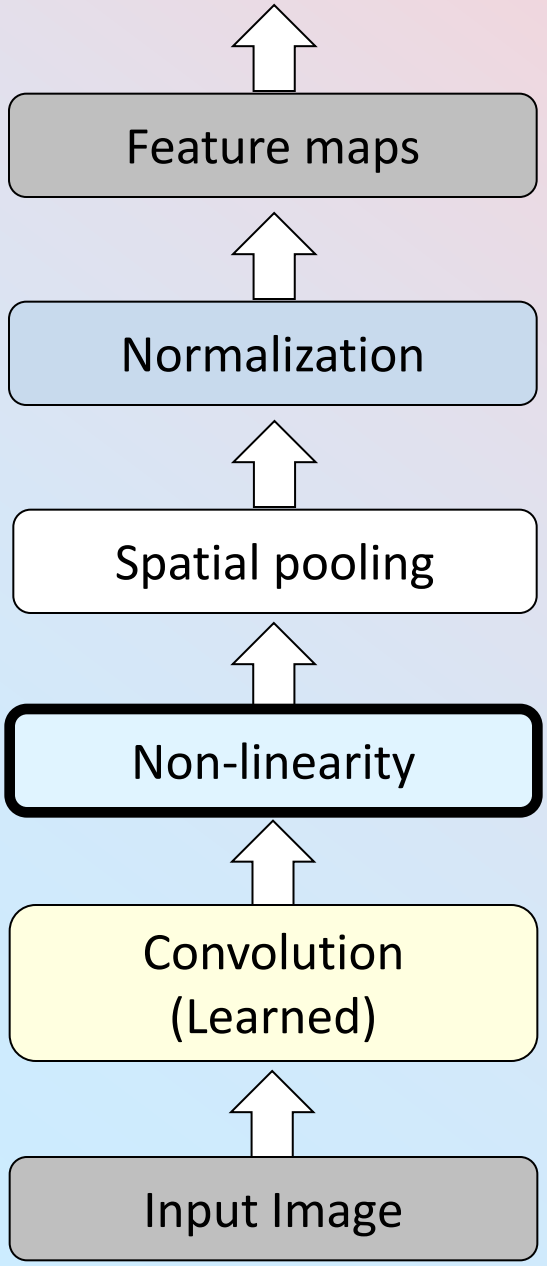
Convolutional Neural Networks



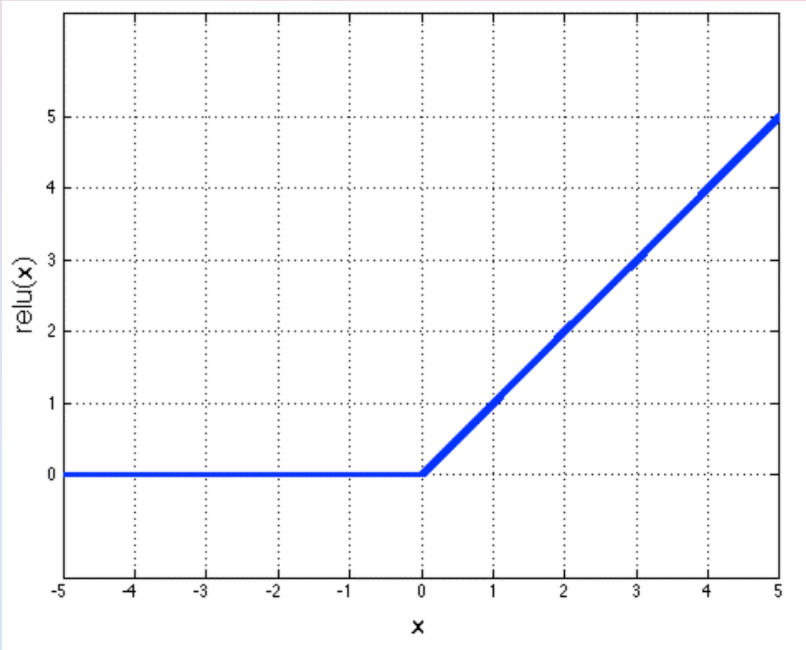


slide credit: S. Lazebnik

Convolutional Neural Networks

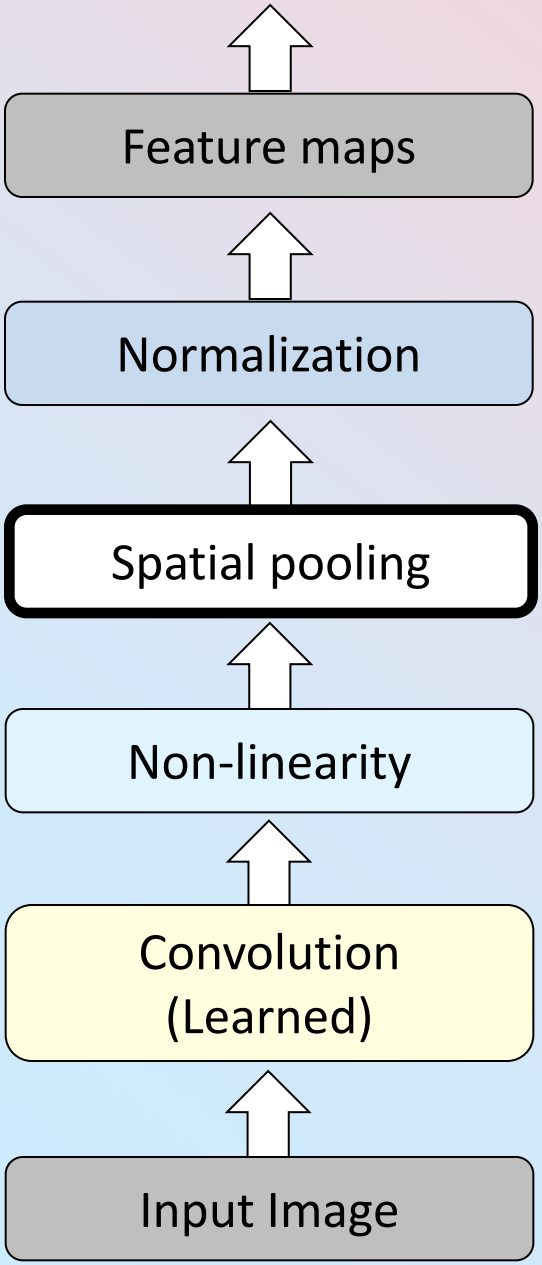


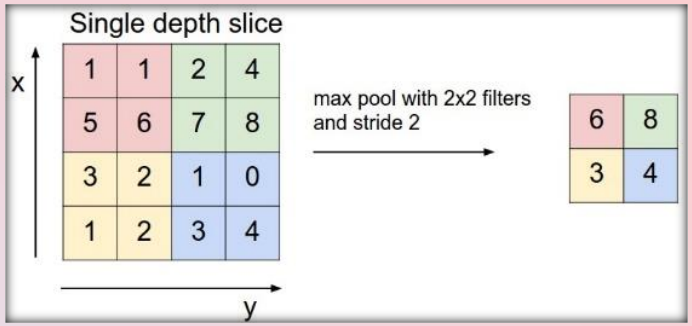
Rectified Linear Unit (ReLU)

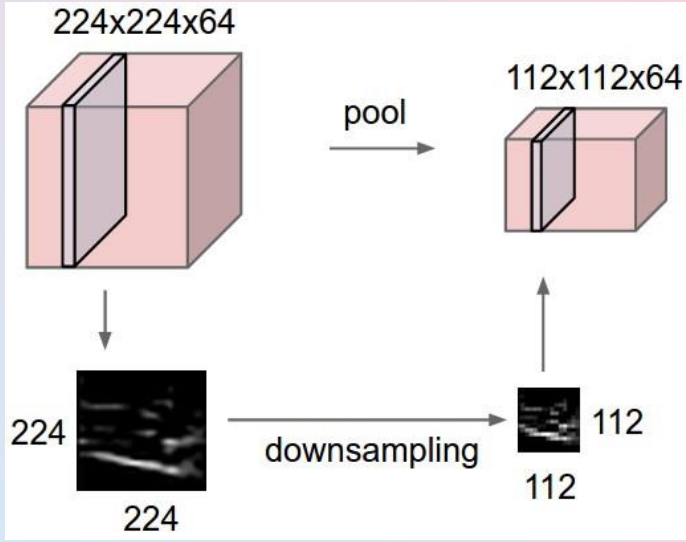


slide credit: S. Lazebnik

Convolutional Neural Networks

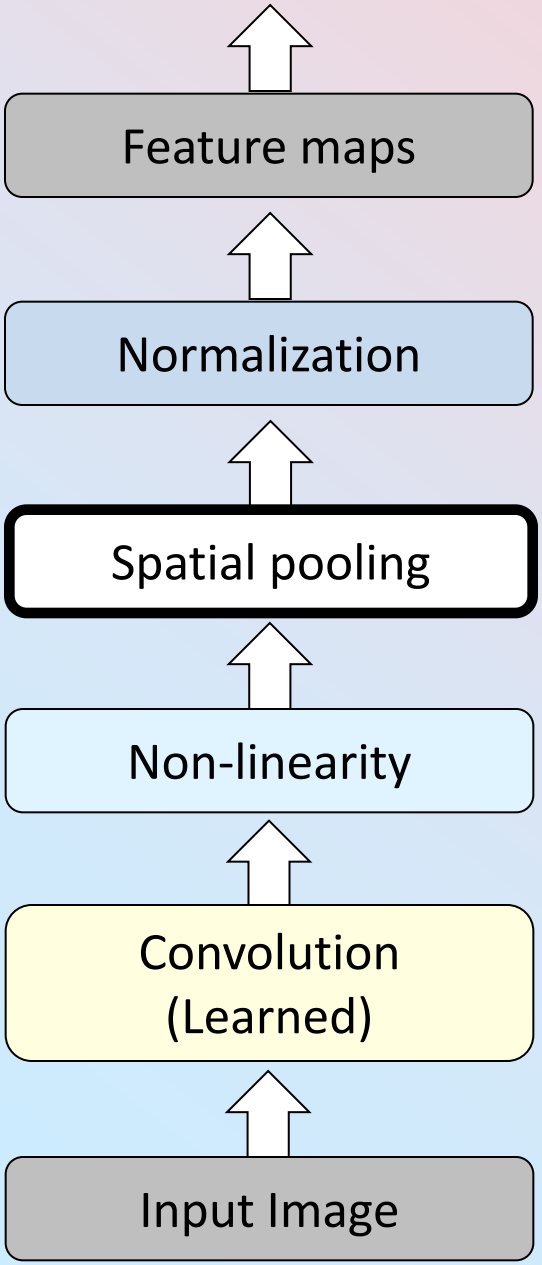


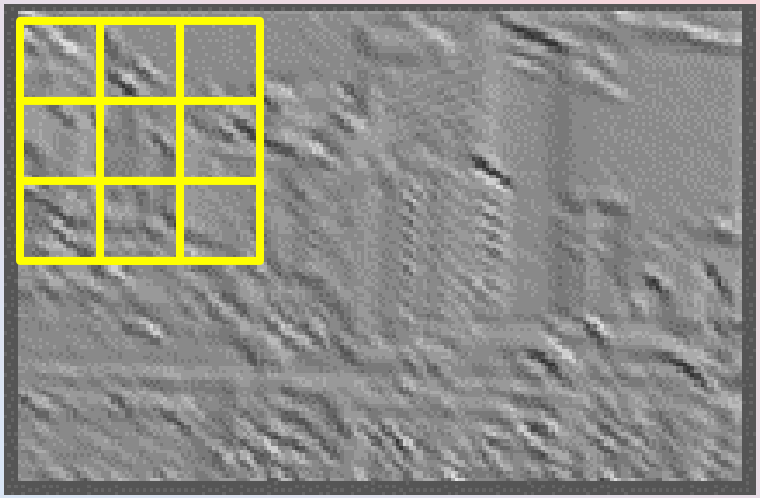




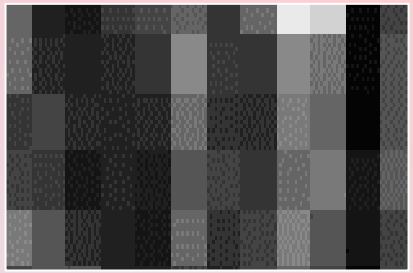
slide credit: S. Lazebnik

Convolutional Neural Networks





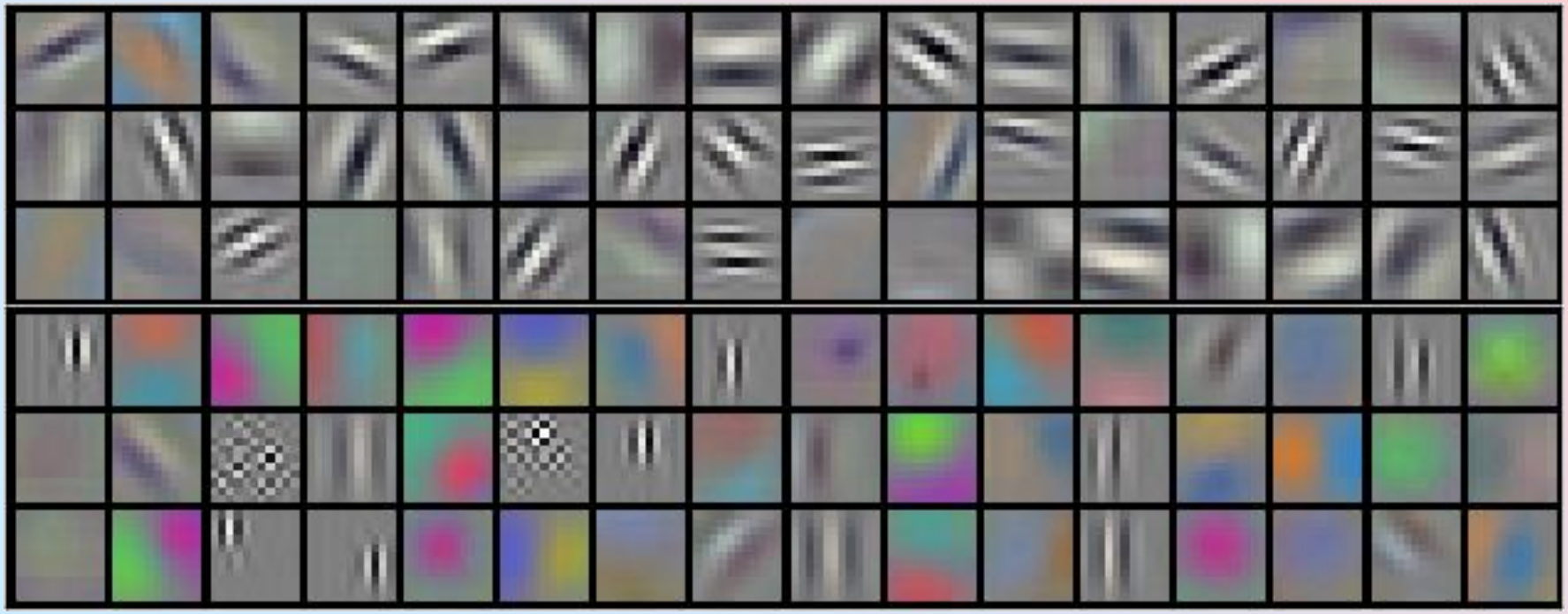
Max pooling



slide credit: S. Lazebnik

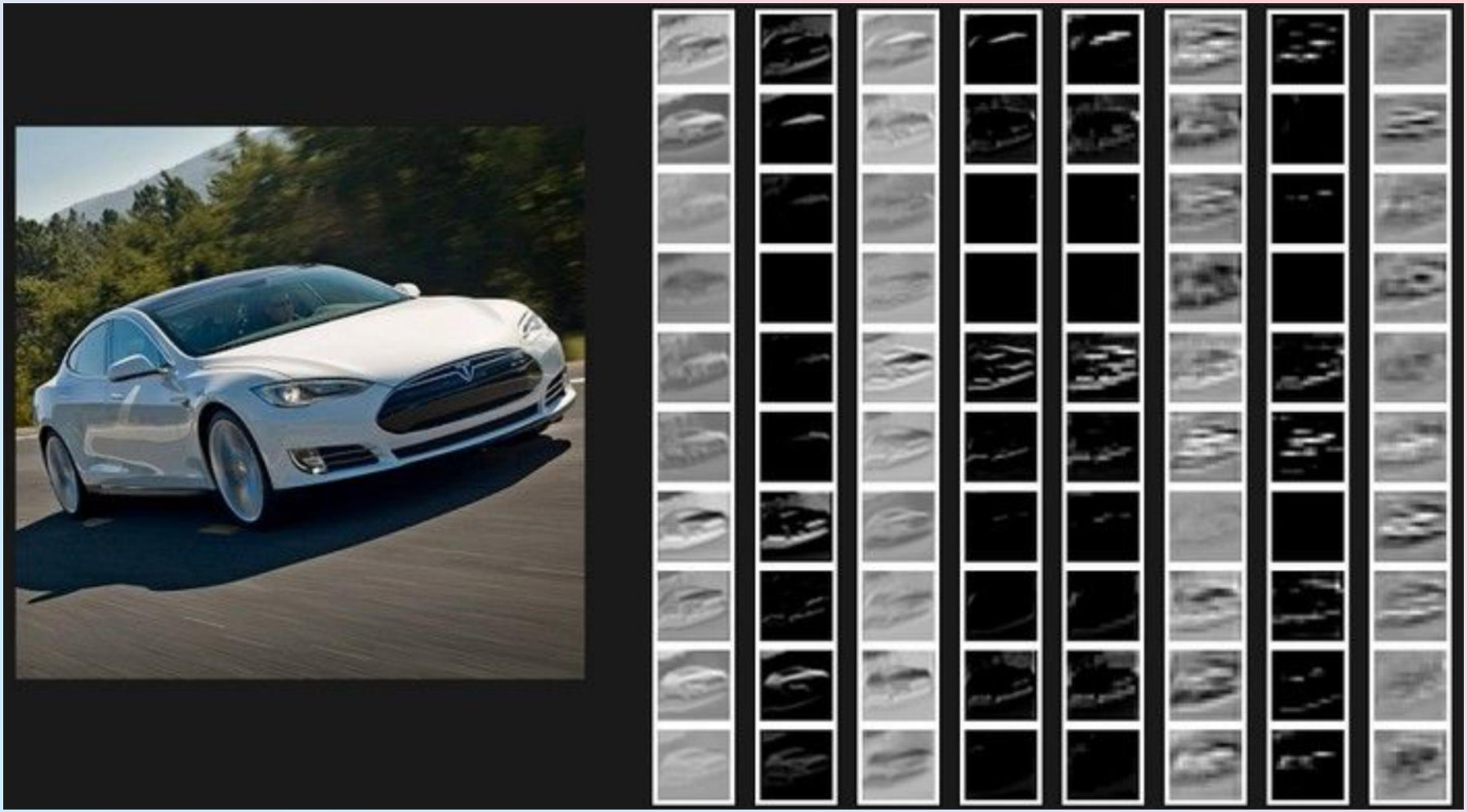
Visualizing what was learned

- What would the learned filters look like?



Visualizing what was learned

- What do the feature maps look like?



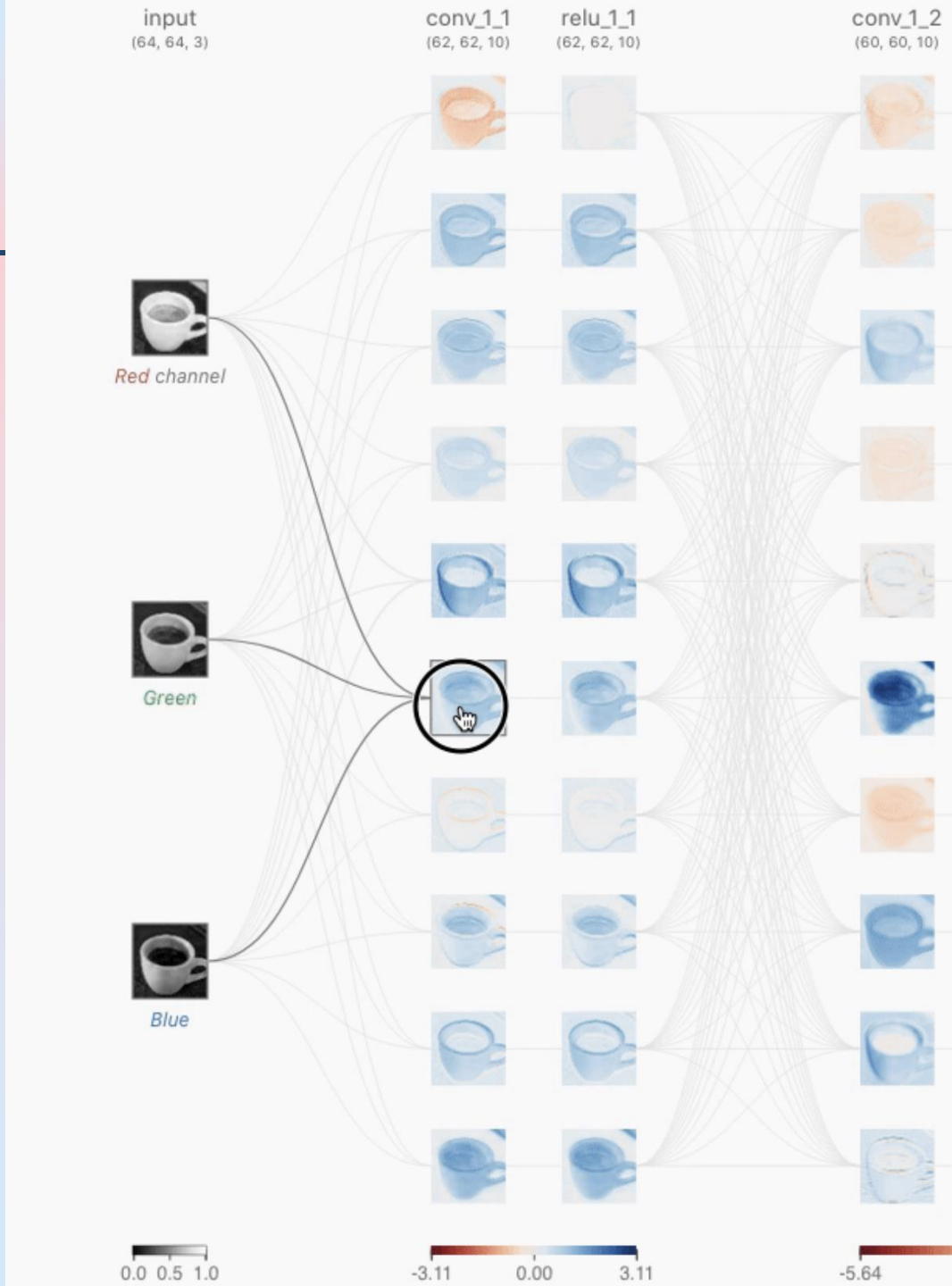
source

The CNN Explainer

Thanks to CS640 alumni Mao Mao, we have a link to the *CNN Explainer*:

https://poloclub.github.io/cnn-explainer/

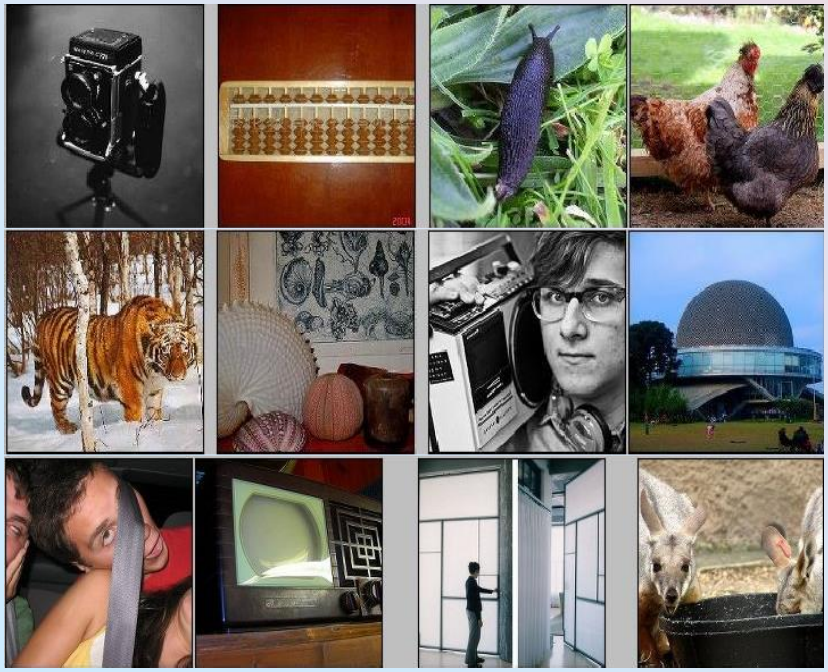
by Jay Wang, Robert Turko, Omar Shaikh, Haekyu Park, Nilaksh Das, Fred Hohman, Minsuk Kahng, and Polo Chau, a result of a research collaboration between Georgia Tech and Oregon State University



ImageNet –
The Data Set that Mattered and Still Matters!

IMAGENET

[Deng et al. CVPR 2009]



- 14 million labeled images
- 20 thousand object classes

- Images collected from the Internet

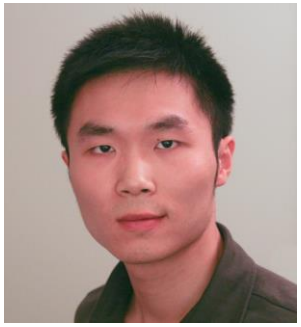
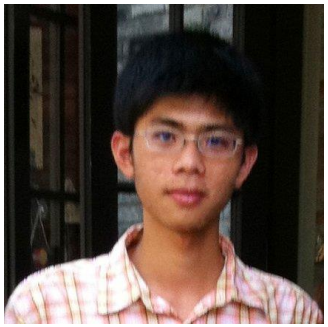
- Human labels obtained by crowdsourcing with Amazon Turk

- Still very important in 2024 because it is widely used for pretraining of “backbone neural nets” of current models

Analysis of Large Scale Visual Recognition

Adapted for BU CS 440/640 by M. Betke

Fei-Fei Li and Olga Russakovsky


Olga Russakovsky, Jia Deng, Zhiheng Huang, Alex Berg, Li Fei-Fei
Detecting avocados to zucchinis: what have we done, and where are we going?
ICCV 2013 **http://image-net.org/challenges/LSVRC/2012/analysis**

Flute

Strawberry

Traffic light

Backpack

Matchstick

Sea lion

Bathing cap

Racket

Large-scale recognition

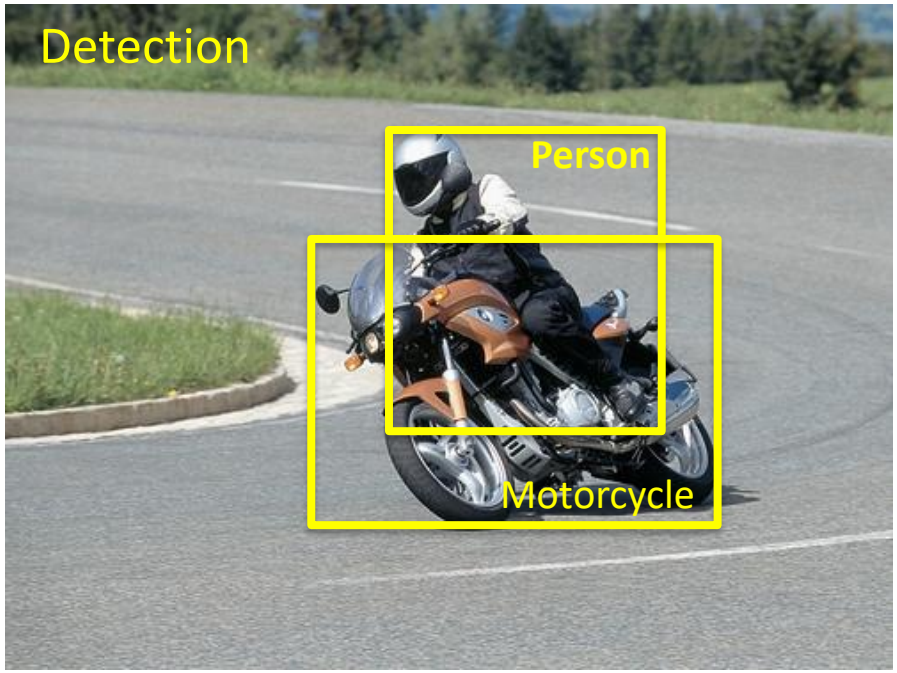
Large-scale recognition

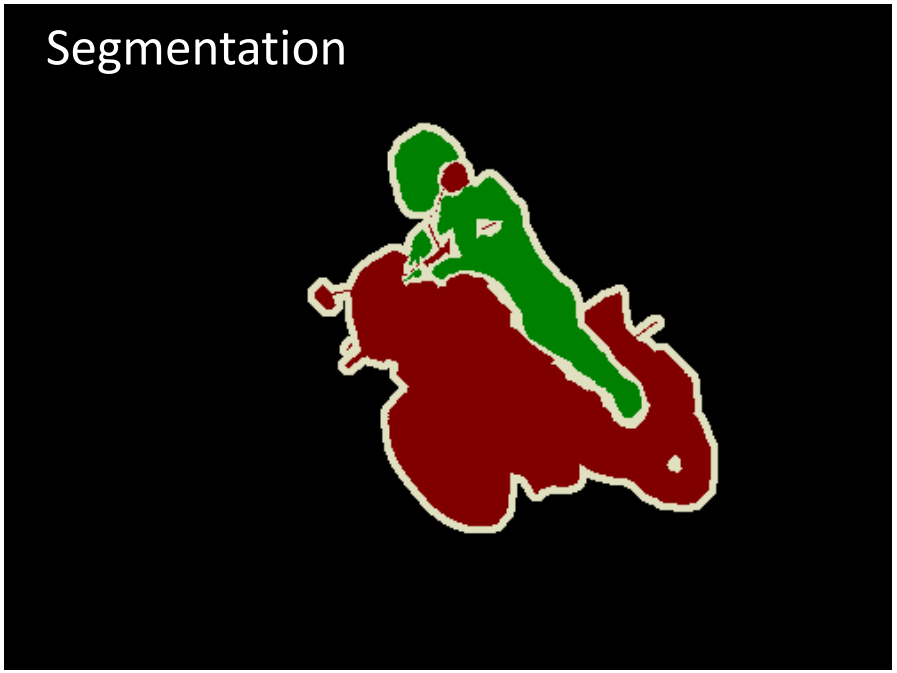
Need benchmark datasets

PASCAL VOC 2005-2012

20 object classes **22,591 images**

Classification: person, motorcycle





Action: riding bicycle

Everingham, Van Gool, Williams, Winn and Zisserman.
The PASCAL Visual Object Classes (VOC) Challenge. IJCV 2010.

IMAGENET Large Scale Visual Recognition Challenge (ILSVRC) 2010-2012

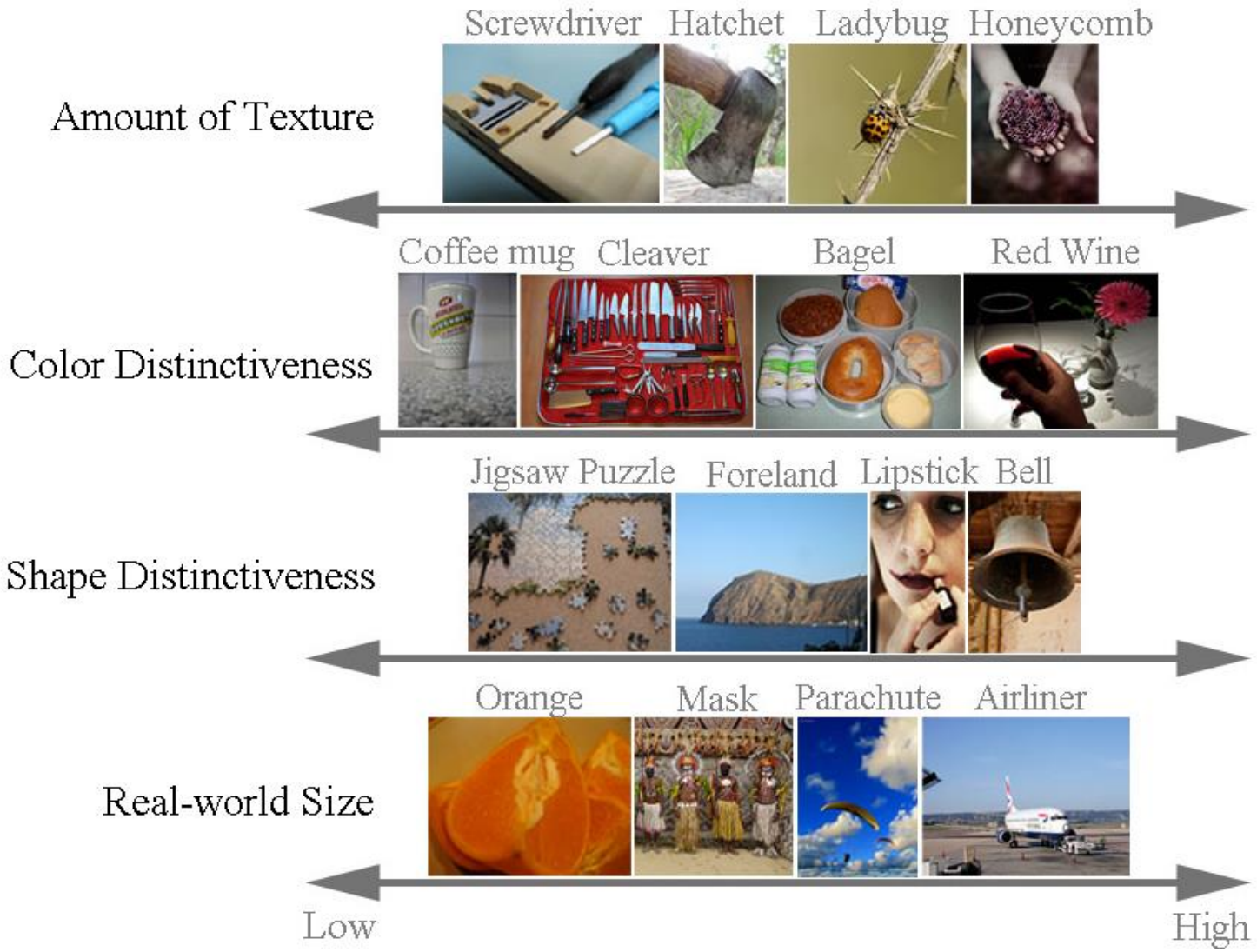
~~20 object classes 22,591 images~~

1000 object classes 1,431,167 images

http://image-net.org/challenges/LSVRC/{2010,2011,2012}

Variety of object classes in ILSVRC

Variety of object classes in ILSVRC



ILSVRC Task 1: Classification

Steel drum

ILSVRC Task 1: Classification

Allowed system output: 5 predictions per image

Goal: Get 1 of the 5 predictions correct

Steel drum

Output:
Scale
T-shirt
Steel drum
Drumstick
Mud turtle

Output:
Scale
T-shirt
Giant panda
Drumstick
Mud turtle

Indicator Function:

1[System output correct on this image] = 1 = 0

ILSVRC Task 1: Classification

Steel drum

Output:
Scale
T-shirt
Steel drum
Drumstick
Mud turtle

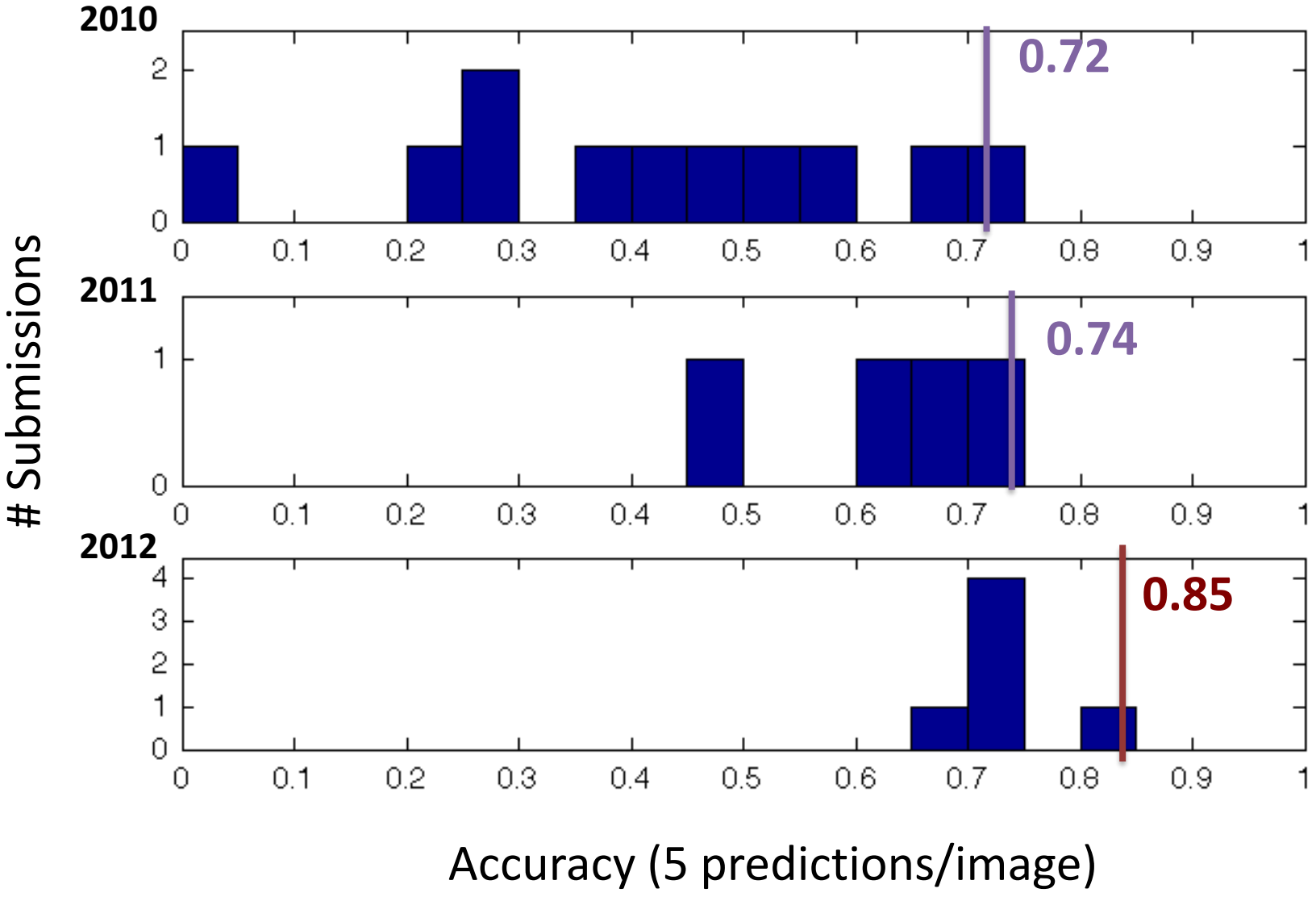
✓

Output:
Scale
T-shirt
Giant panda
Drumstick
Mud turtle

✗

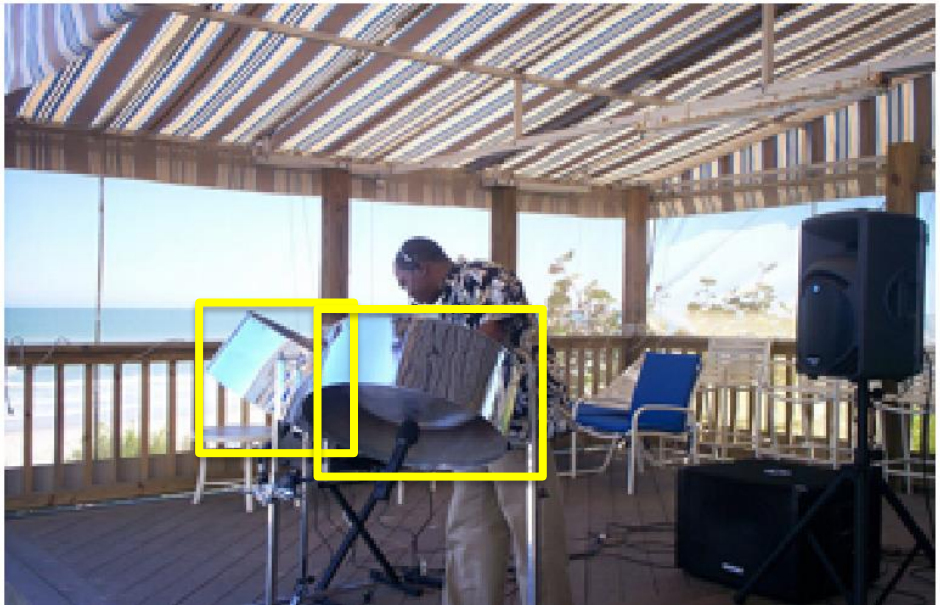
$$\text{Accuracy} = \frac{1}{100{,}000} \sum_{\substack{100{,}000 \\ \text{images}}} 1[\text{correct on image i}]$$

ILSVRC Task 1: Classification



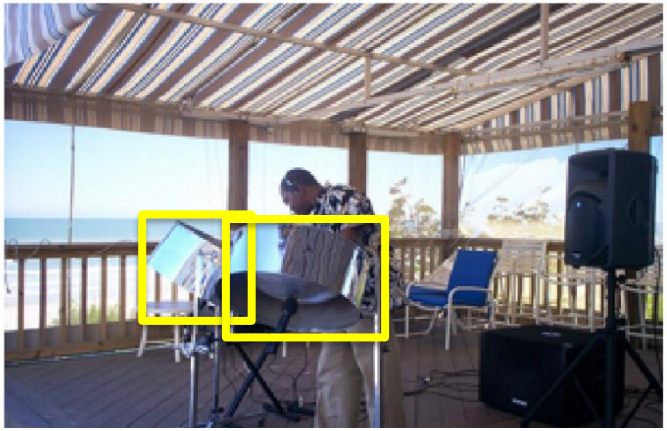
ILSVRC Task 2: Classification + Localization

Steel drum

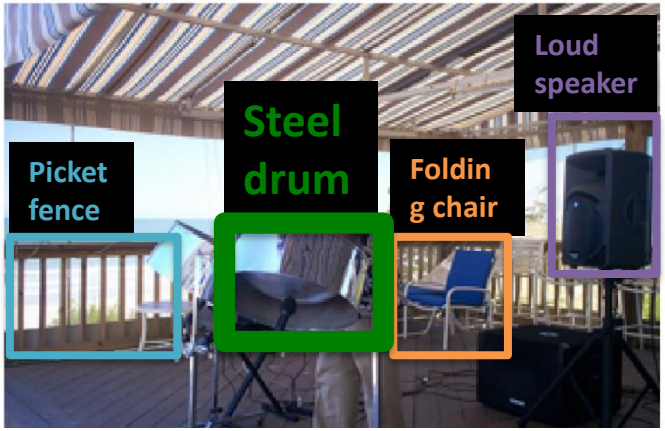


ILSVRC Task 2: Classification + Localization

Steel drum



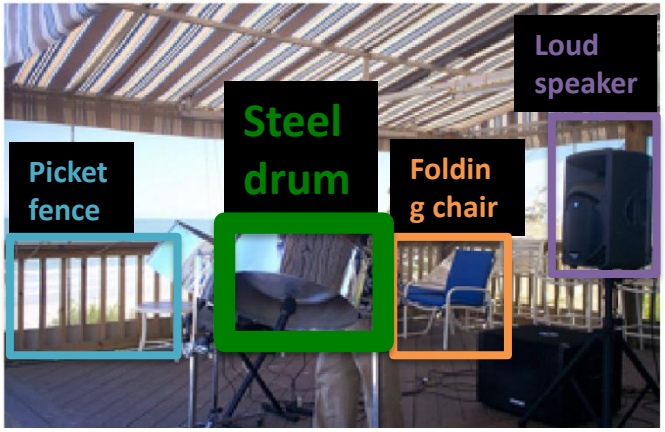
Output



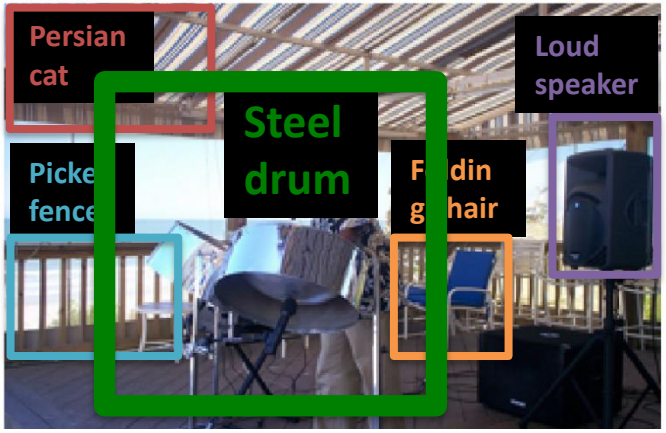
ILSVRC Task 2: Classification + Localization

Steel drum

Output



Output (bad localization)

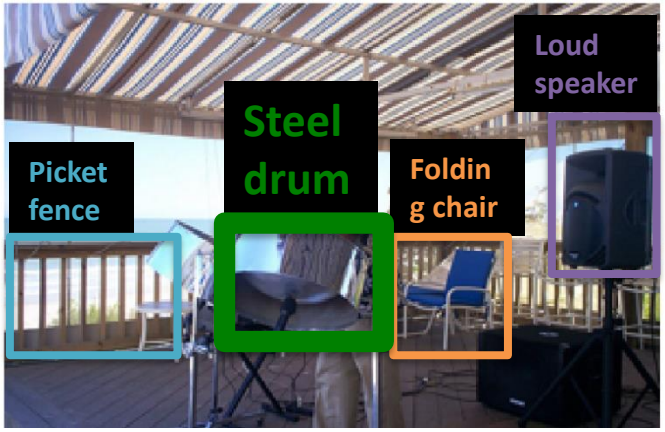


Output (bad classification)

ILSVRC Task 2: Classification + Localization

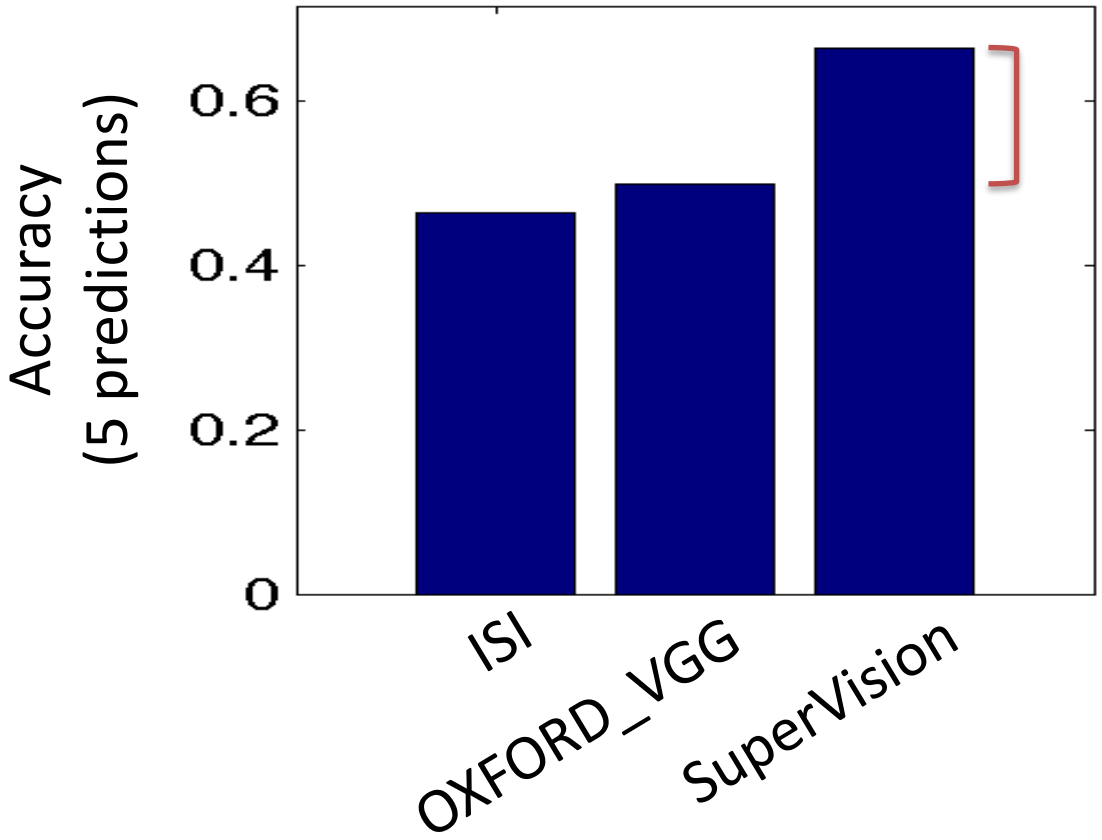
Steel drum

Output



$$\text{Accuracy} = \frac{1}{100{,}000} \sum_{\substack{100{,}000 \\ \text{images}}} 1[\text{correct on image i}]$$

ILSVRC Task 2: Classification + Localization



ISI=Uni. Tokyo Team

VGG=Uni. Oxford Team

SuperVision =
University of Toronto Team
Led by
Geoffrey Hinton,
Turing Award and Nobel
Price Winner

What happens under the hood?

Preliminaries:

- ILSVRC-500 (2012) dataset
- Leading algorithms

What happens under the hood on classification+localization?

- A closer look at small objects
- A closer look at textured objects

Olga Russakovsky, Jia Deng, Zhiheng Huang, Alex Berg, Li Fei-Fei
Detecting avocados to zucchinis: what have we done, and where are we going?
ICCV 2013 http://image-net.org/challenges/LSVRC/2012/analysis

ILSVRC (2012)

1000 object classes

Easy to localize

Hard to localize

ILSVRC-500 (2012)

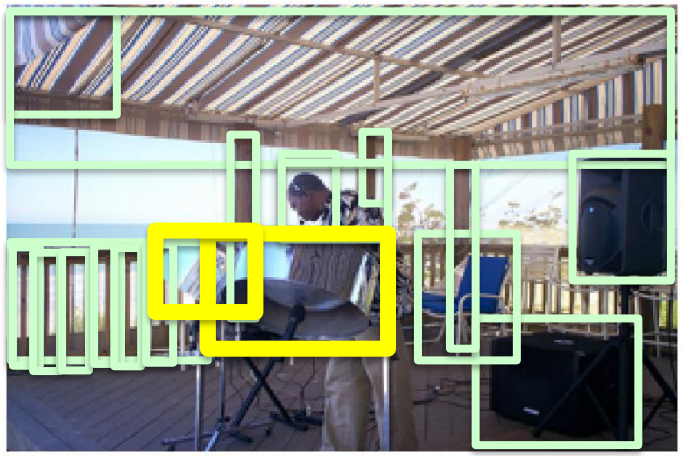
ILSVRC-500 (2012)

Object scale (fraction of image area occupied by target object)

ILSVRC-500 (2012)	500 object categories	25.3%
PASCAL VOC (2012)	20 object categories	25.2%

Level of clutter

Steel drum



- Generate candidate object regions using method of
  Selective Search for Object Detection
  vanDeSande et al. ICCV 2011
- Filter out regions inside object
- Count regions

ILSVRC-500 (2012)	500 object categories	128 ± 35
PASCAL VOC (2012)	20 object categories	130 ± 29

SuperVision = AlexNet

Alex Krizhevsky, Ilya Sutskever, Geoffrey Hinton (Krizhevsky NIPS12)

Image classification: Deep convolutional neural networks

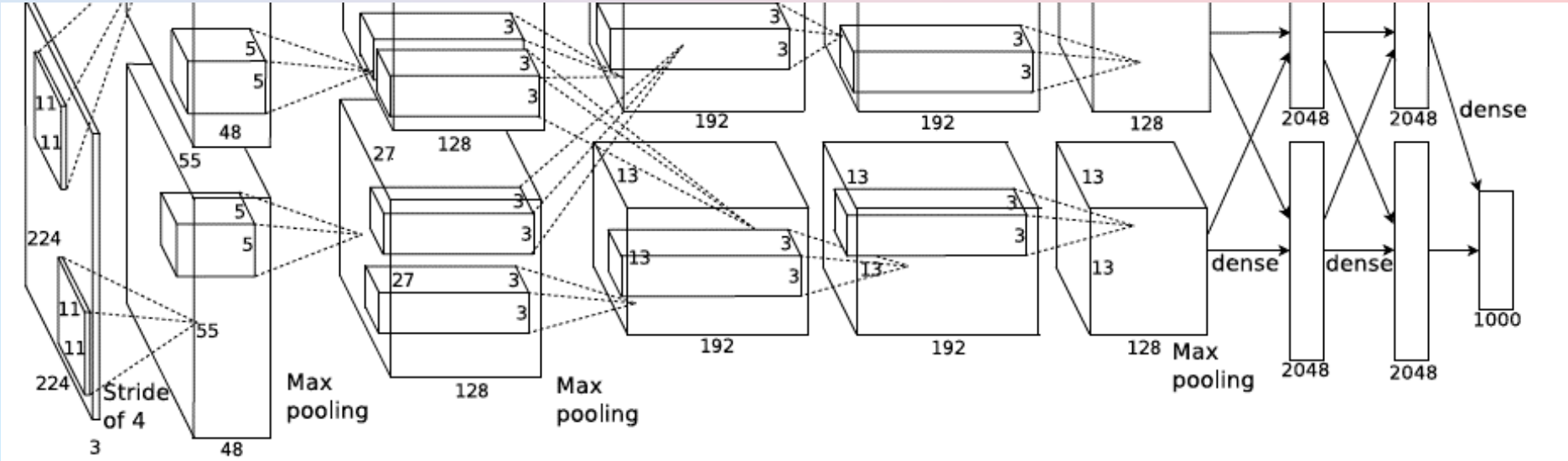
- 7 hidden "weight" layers, 650K neurons, 60M parameters, 630M connections
- Rectified Linear Units, max pooling, dropout trick
- Randomly extracted 224x224 patches for more data
- Trained with Stochastic Gradient Descent on two GPUs for a week, fully supervised (50x speed-up over CPU)

Localization: Regression on (x,y,w,h)

http://www.cs.toronto.edu/~fritz/absps/imagenet.pdf

AlexNet

- Similar to the model proposed by LeCun in 1998 but:
  - Larger model (7 hidden layers, 650,000 units, 60,000,000 params)
  - More data (10^6 vs. 10^3 images)



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

Details of the Oxford VGG

This is **not** the neural net VGG but uses traditional computer vision techniques!

Karen Simonyan, Yusuf Aytar, Andrea Vedaldi, Andrew Zisserman

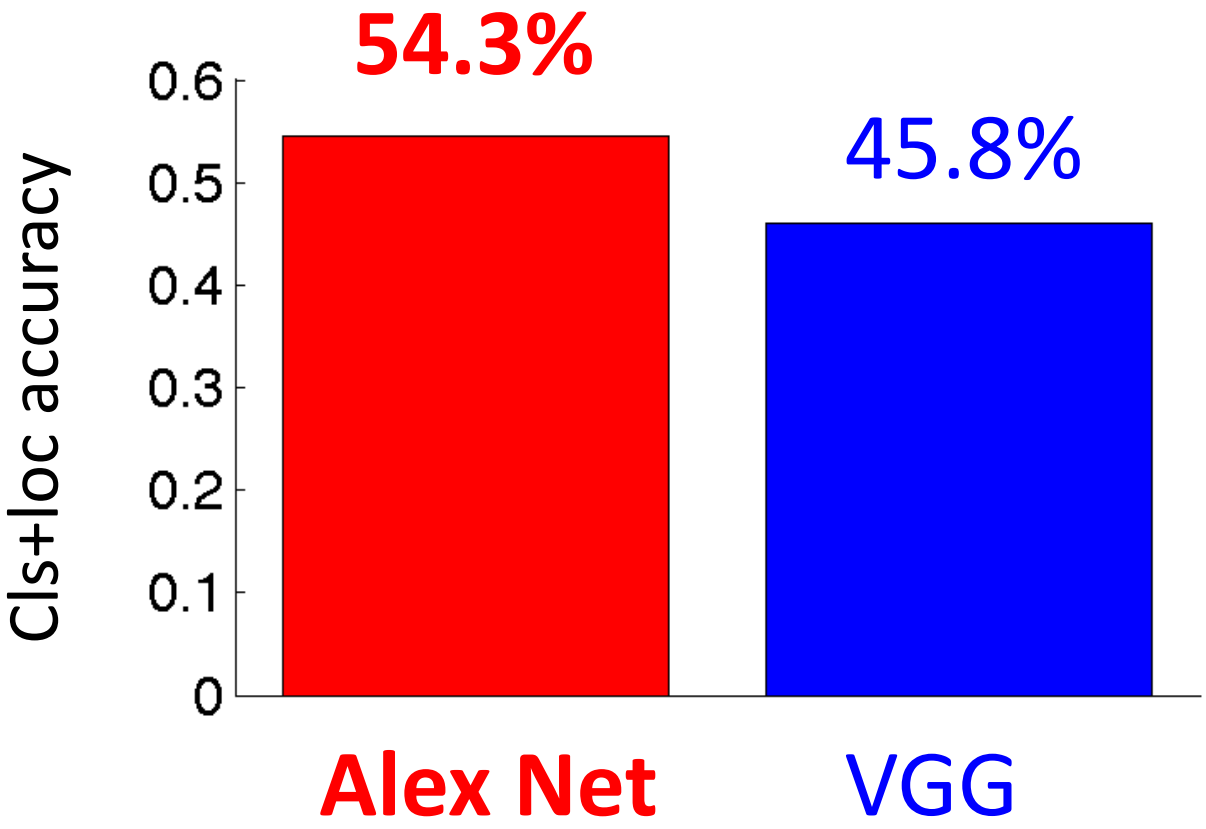
Image classification: Fisher vector + linear SVM (Sanchez CVPR11)

- Root-SIFT (Arandjelovic CVPR12), color statistics, augmentation with patch location (x,y) (Sanchez PRL12)
- Fisher vectors: 1024 Gaussians, 135K dimensions
- Product quantization to compress
- Semi-supervised learning to find additional bounding boxes
- 1000 one-vs-rest SVM trained with Pegasos SGD
  - 135M parameters!

Localization: Deformable part-based models (Felzenszwalb PAMI10), without parts (root-only)

https://image-net.org/static_files/files/oxford_vgg.pdf

Results on ILSVRC-500



Preliminaries:

- ILSVRC-500 (2012) dataset – similar to PASCAL
- Leading algorithms: Alex Net and VGG

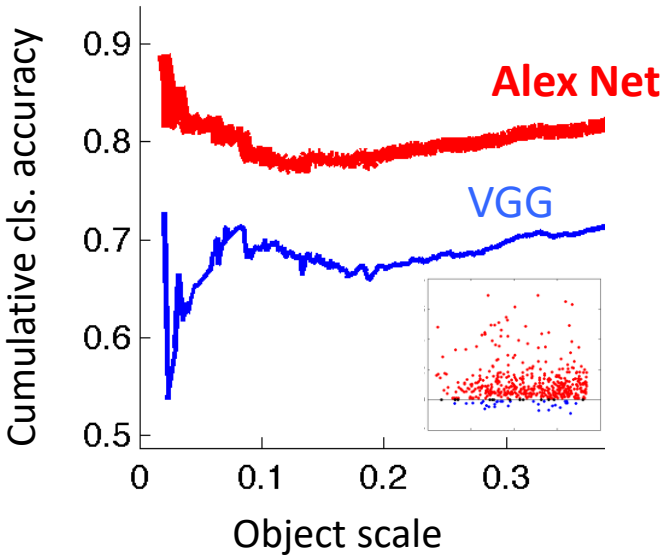
What happens under the hood on classification+localization?

- Alex Net always great at classification, but VGG does better than Alex Net localizing small objects
- A closer look at textured objects

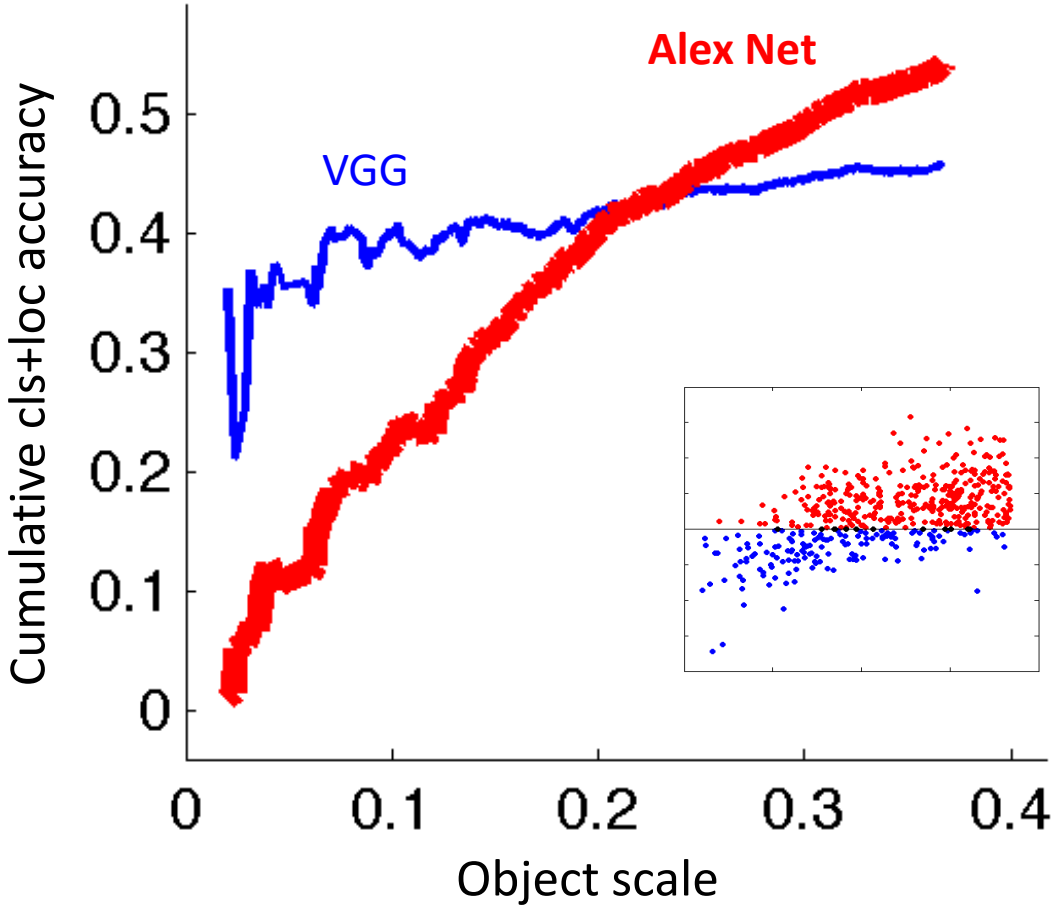
Olga Russakovsky, Jia Deng, Zhiheng Huang, Alex Berg, Li Fei-Fei
Detecting avocados to zucchinis: what have we done, and where are we going?
ICCV 2013 http://image-net.org/challenges/LSVRC/2012/analysis

Cumulative accuracy across scales

Classification-only

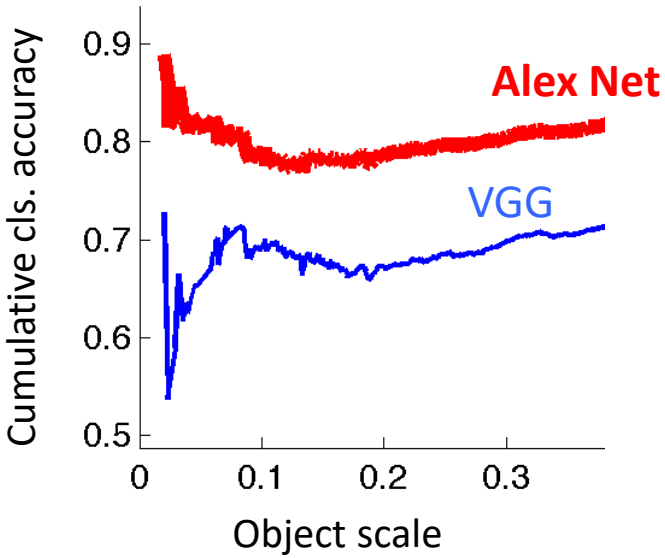


Classification+Localization

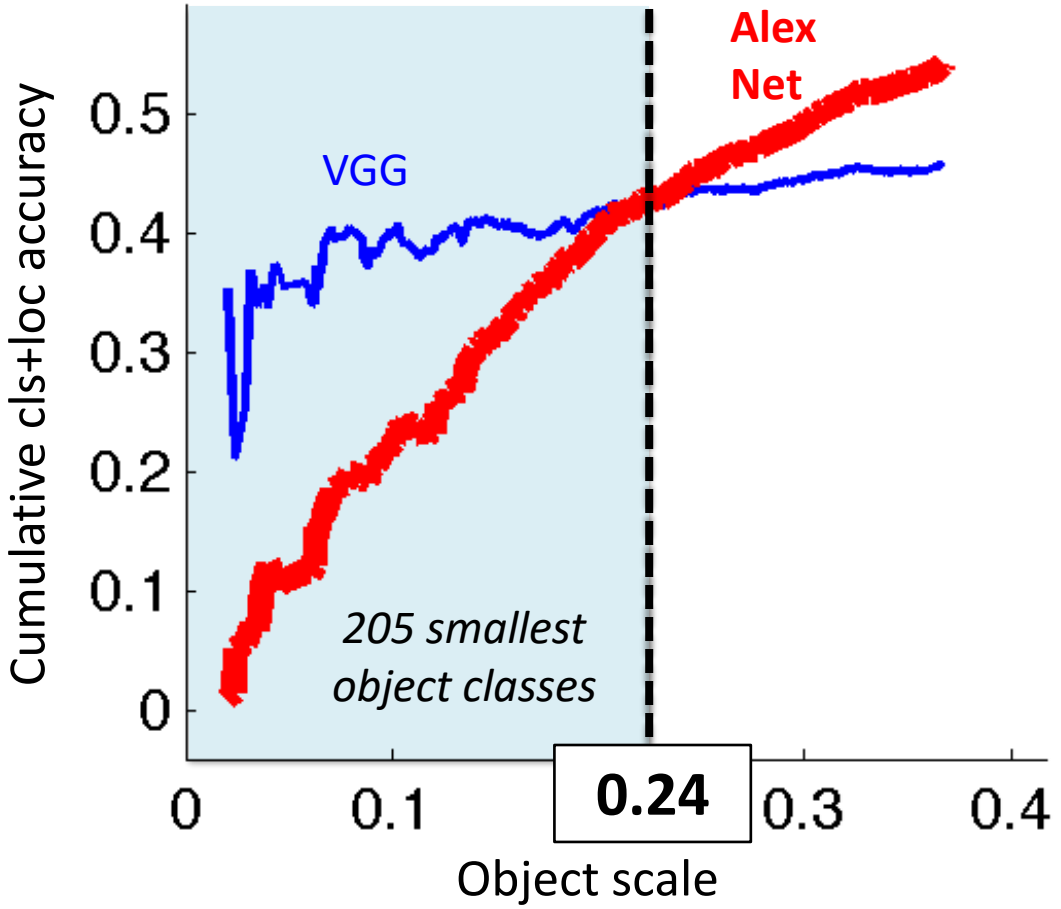


Cumulative accuracy across scales

Classification-only



Classification+Localization



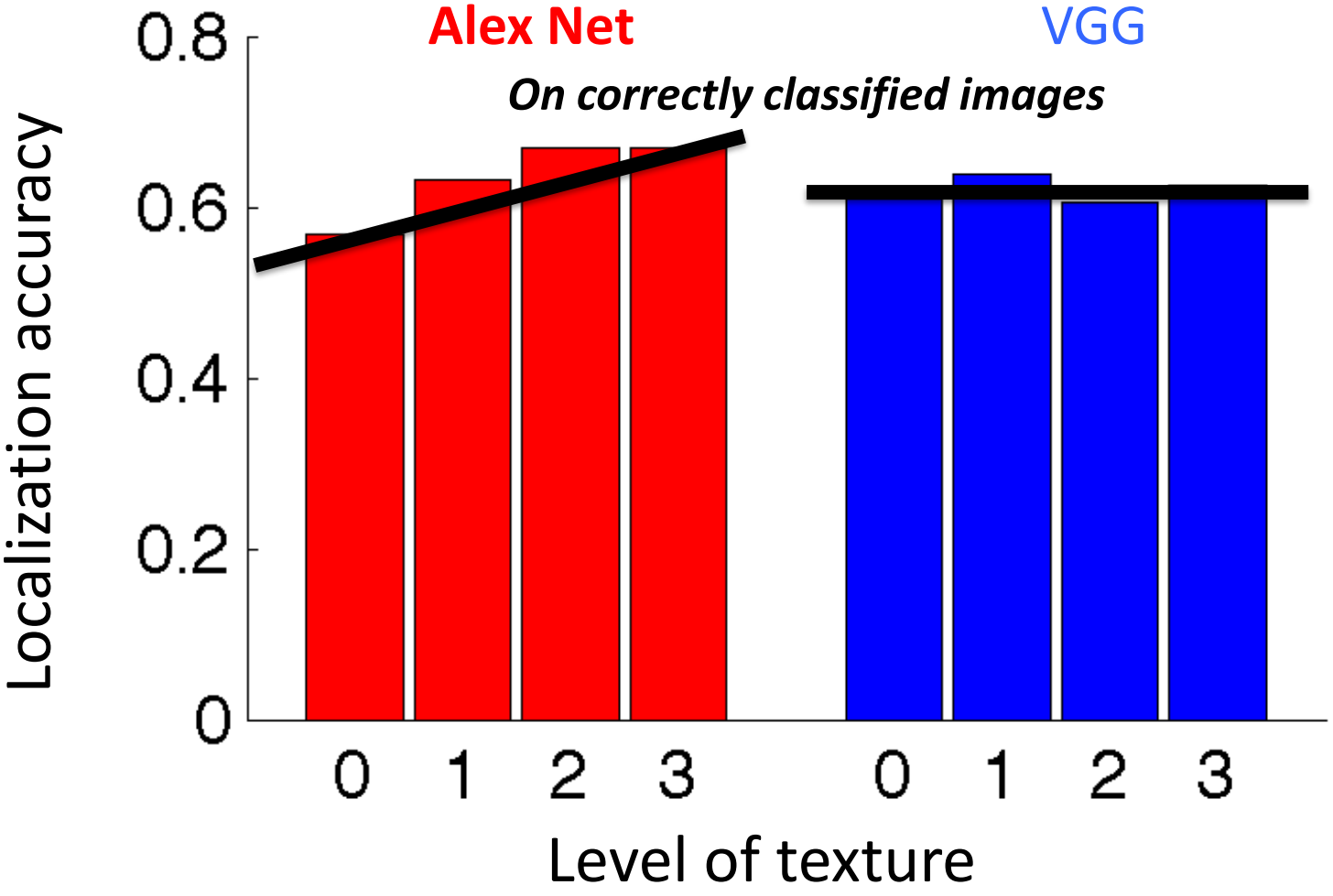
Textured objects (ILSVRC-500)

Textured objects (ILSVRC-500)

	No texture	Low texture	Medium texture	High texture
# classes	116	189	143	52
Object scale	20.8%	23.7%	23.5%	25.0%

Localizing textured objects

(416 classes, same average object scale at each level of texture)

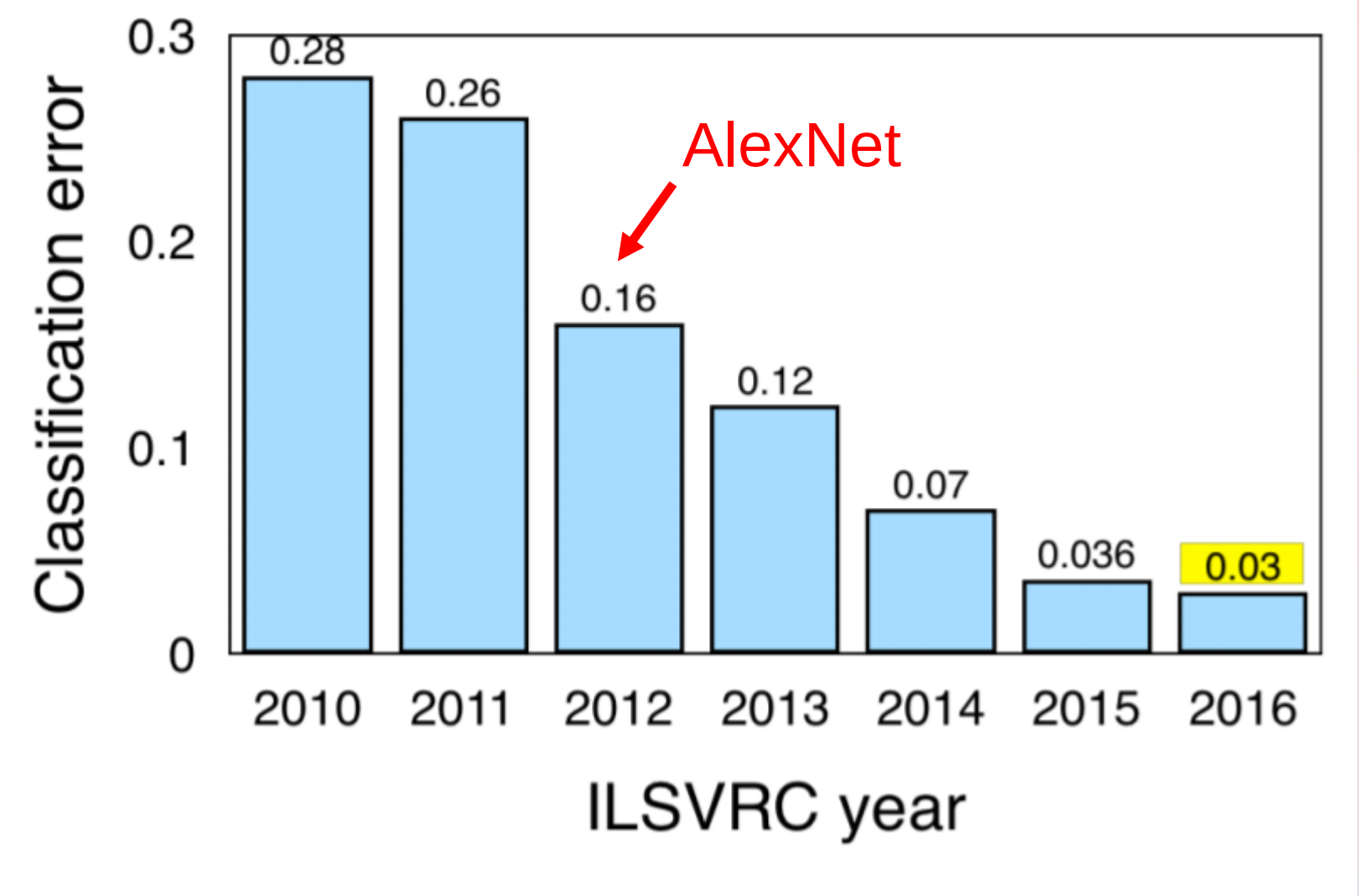


Conclusions on analysis of classification+localization results

- Alex Net always great at classification, but VGG does better than Alex Net localizing small objects
- Textured objects: VGG broadly successful. Alex Net better at higher textures, worse at smaller.

Olga Russakovsky, Jia Deng, Zhiheng Huang, Alex Berg, Li Fei-Fei
Detecting avocados to zucchinis: what have we done, and where are we going?
ICCV 2013 http://image-net.org/challenges/LSVRC/2012/analysis

ImageNet Classification Challenge



http://image-net.org/challenges/talks/2016/ILSVRC2016_10_09_clsloc.pdf

Recap of NN-based Computer Vision

- Neural networks
  - View of neural networks as learning hierarchy of features
- Convolutional neural networks
  - Architecture of network accounts for image structure
  - “End-to-end” recognition from pixels
  - Together with large labeled datasets and lots of computation → major success on benchmark ImageNet, i.e., object classification and localization

Learning Objectives for this Lecture

Computer Science

- Understand template matching with image pyramids
- Understand CNNs as a learning hierarchy of features
- Learn about early CNN used in computer vision: LeCun's work on recognizing handwritten numbers
- Understand CNN concepts, e.g., convolution layers, fully connected (dense) layers, non-linearity (ReLU), pooling (downsampling)
- Learn about breakthrough dataset ImageNet