A Recent History of Deep Learning and Object Recognition

Larry Zitnick

facebook
Artificial Intelligence Research

Motivation

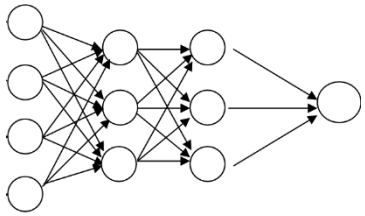
History of Deep Learning applied to classification and detection.

Where are we going?

Goals

20 min: History deep learning

100 min: Current status (RCNN++)

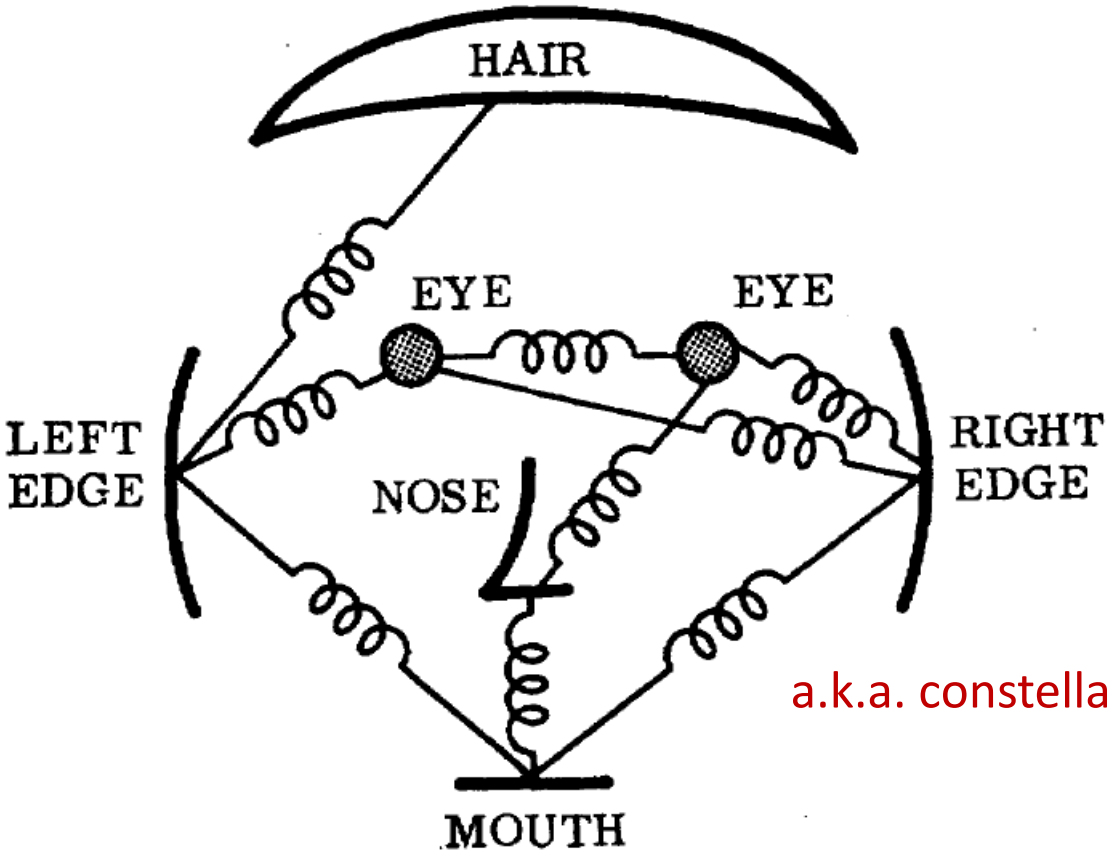


Overview of the last 4-5 years.

15 min: What's next?

After we recognize everything, what next?

1973

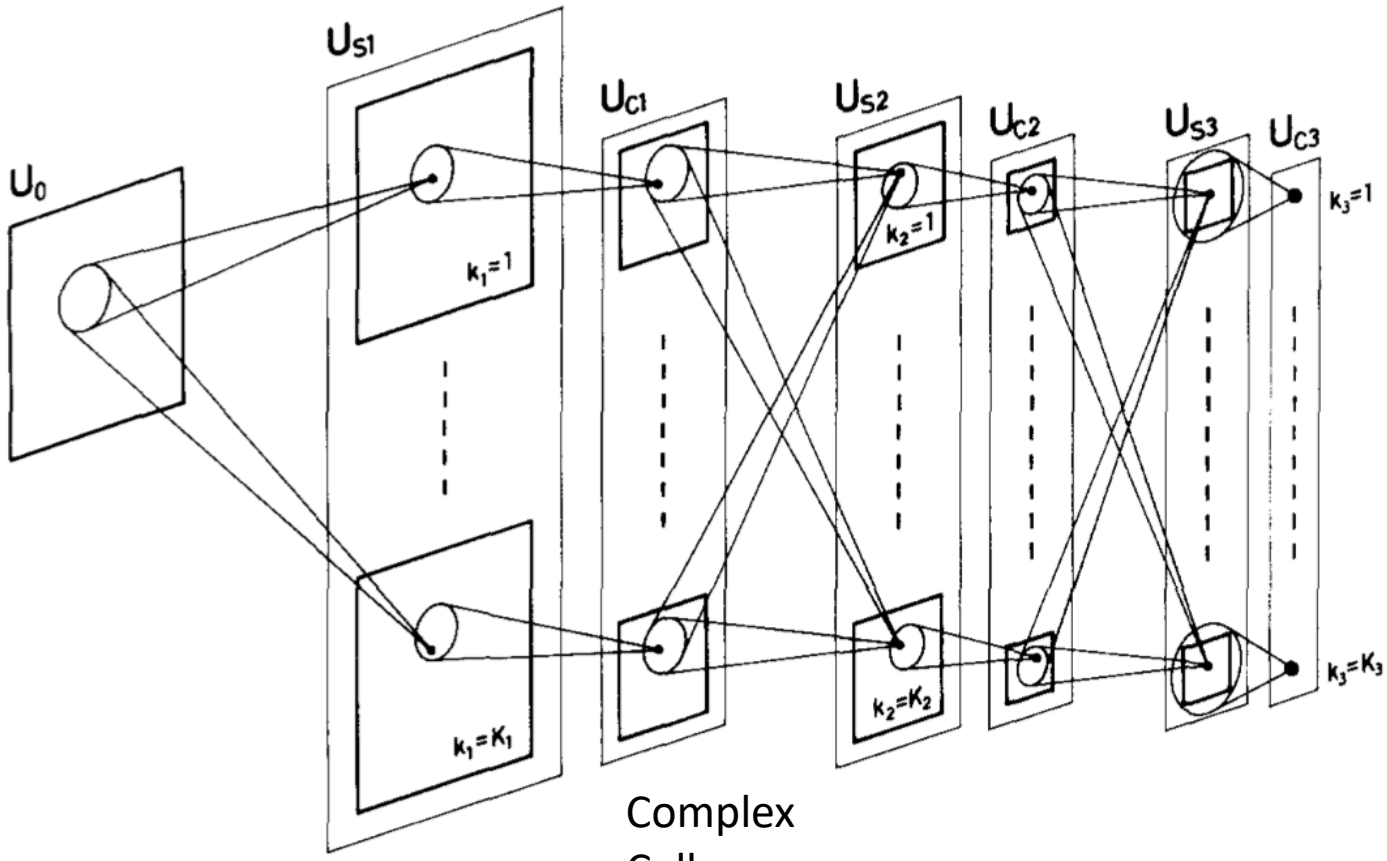


a.k.a. constellation model

The representation and matching of pictorial structures, Fischler and Elschlager, 1973

1980 Neocognitron

Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition **Unaffected by Shift in Position**.

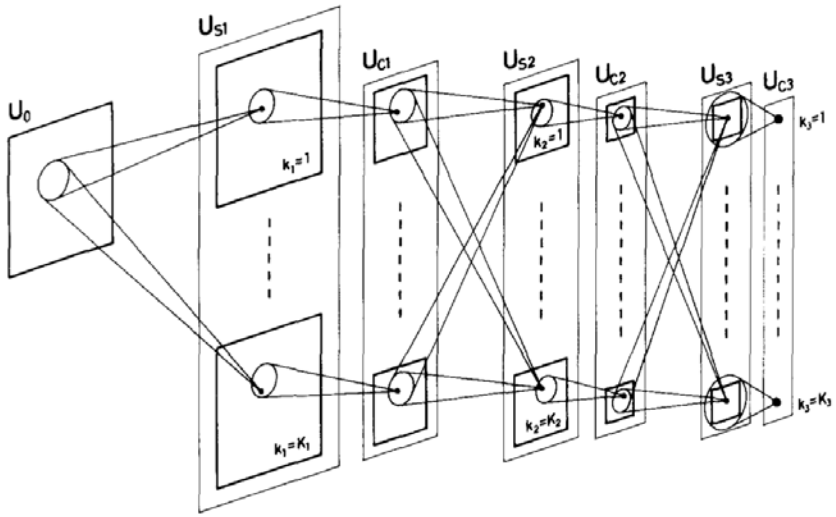


Fukushima. 1980

Activation function

$$\varphi[x] = \begin{cases} x & x \geqq 0 \\ 0 & x < 0 \end{cases}$$

Half-wave rectifier
(ReLU in 1980!)



What's
Missing?

1986 Backpropagation

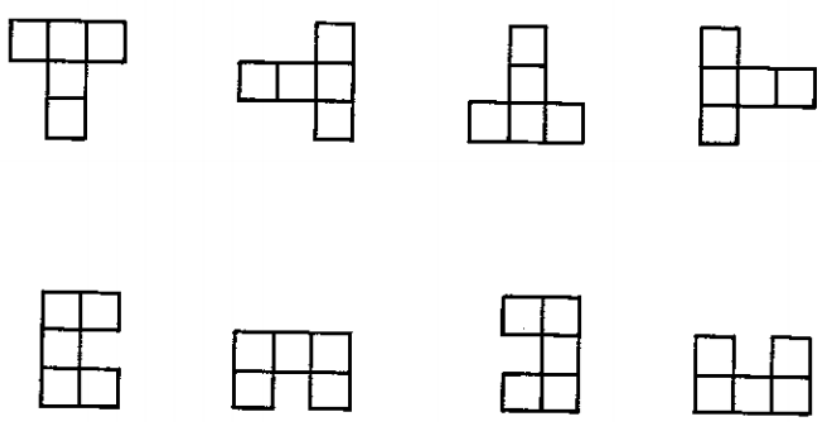


FIGURE 13. The stimulus set for the T-C problem. The set consists of a block T and a block C in each of four orientations. One of the eight patterns is presented on each trial.

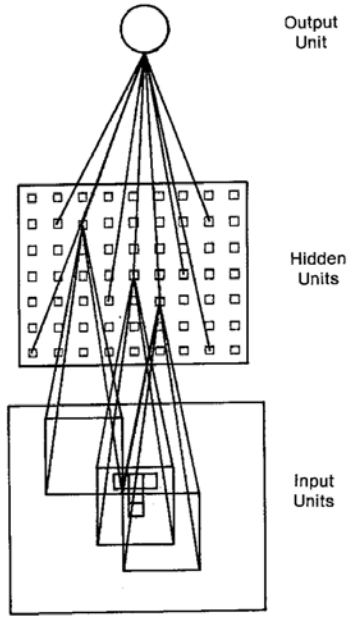
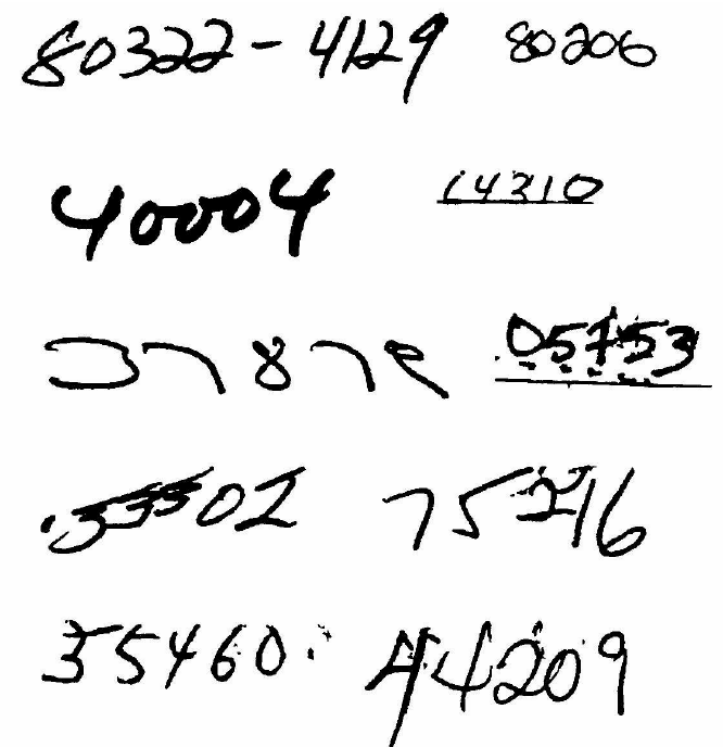


FIGURE 14. The network for solving the T-C problem. See text for explanation.

Learning Internal Representations by Error Propagation,
Rumelhart, Hinton, Williams. 1986.

1989

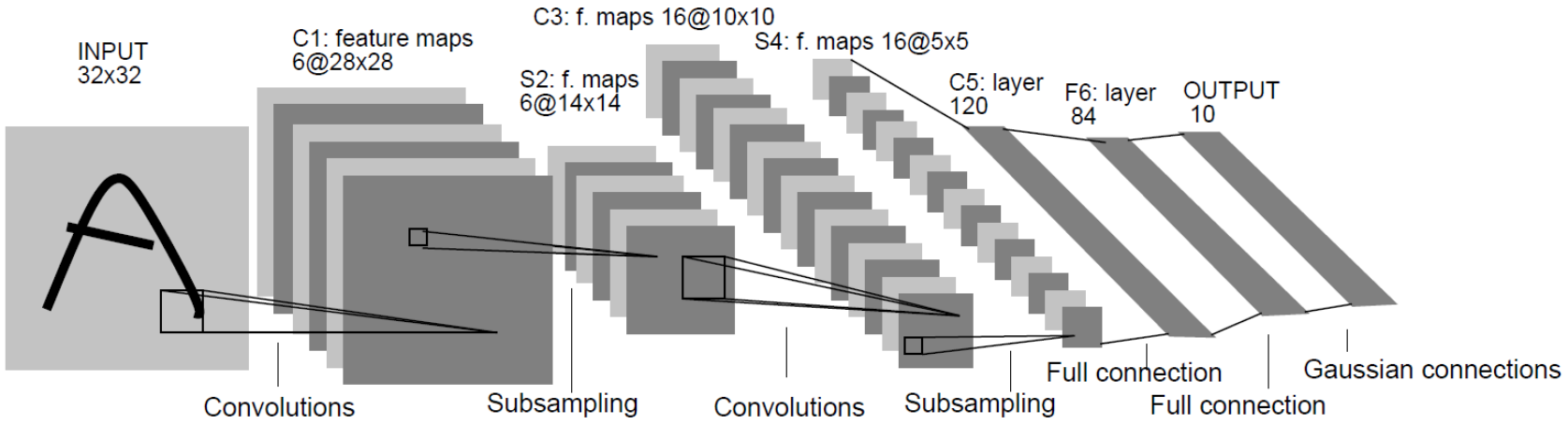
MNIST



Zip codes

Backpropagation applied to handwritten zip code recognition,
Lecun et al., 1989

1989



Backpropagation applied to handwritten zip code recognition,
Lecun et al., 1989

1995 Navlab 5

98% of the way across USA!
~60 MPH

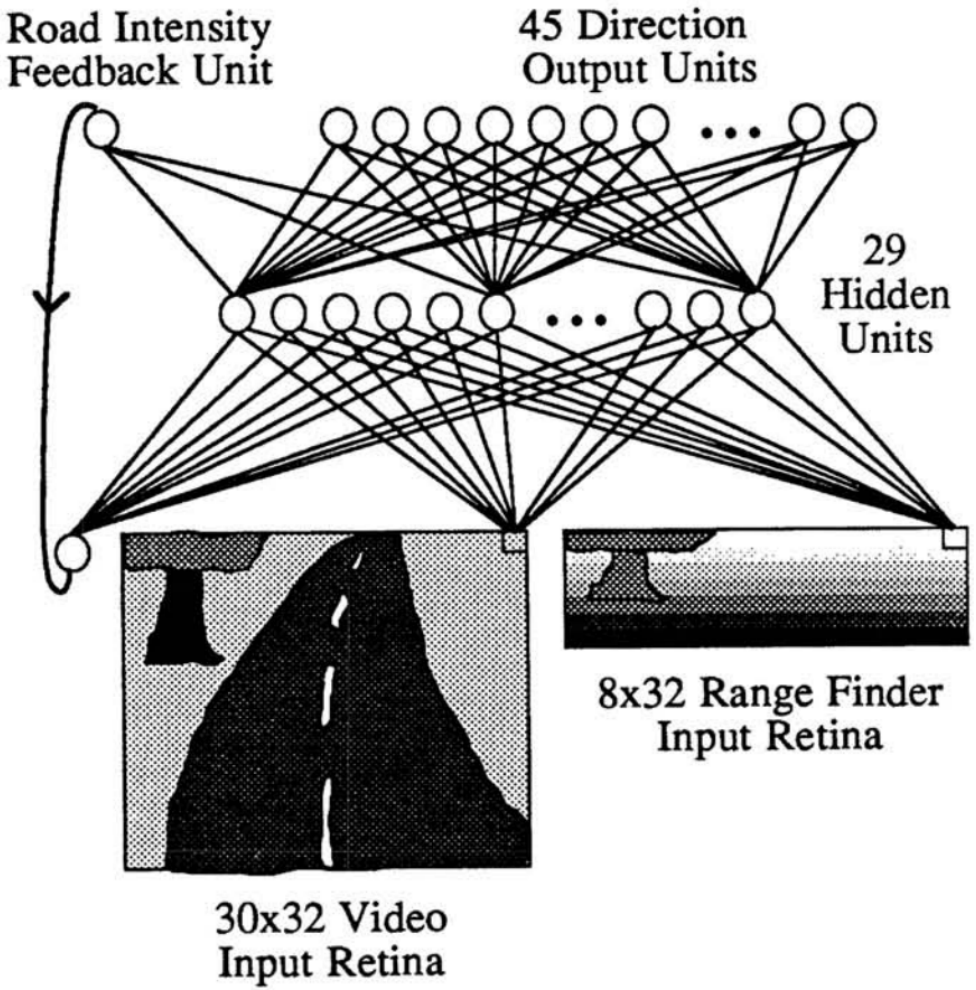
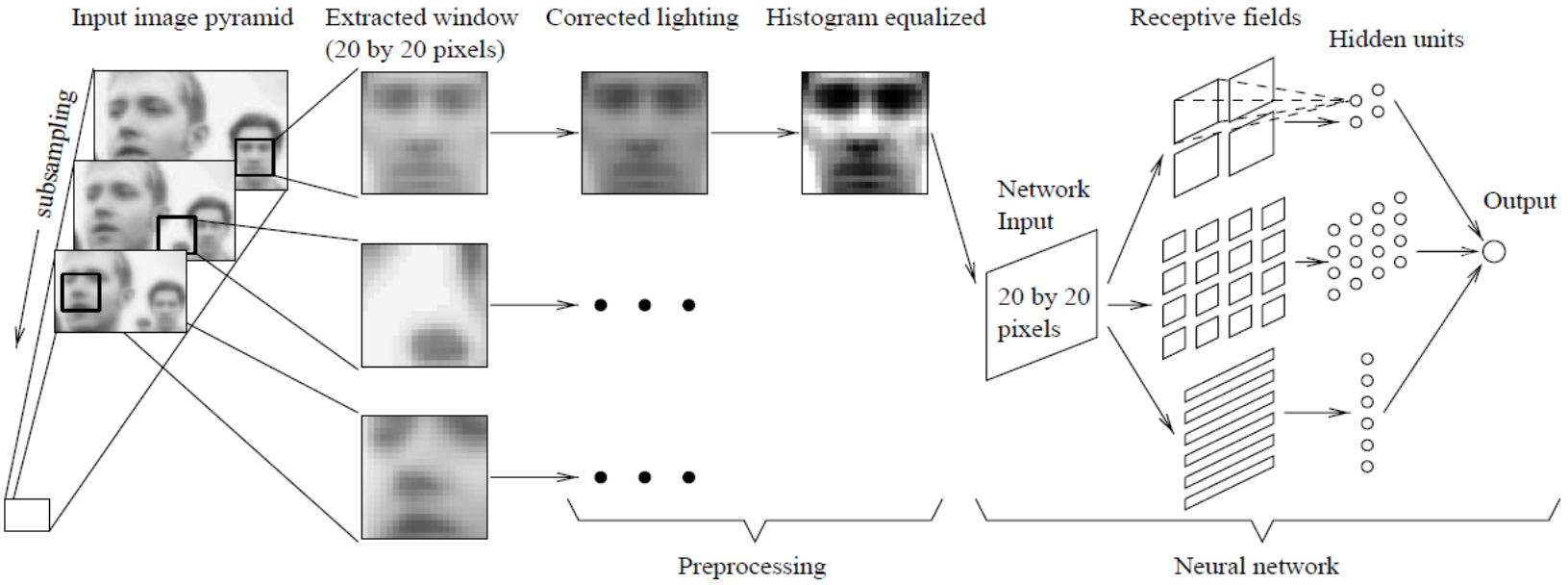


Figure 1: ALVINN Architecture

Alvinn: An autonomous land vehicle in a neural network,
Dean Pomerleau

1998

Faces



Neural Network-Based Face Detection,
Rowley at al., PAMI 1998

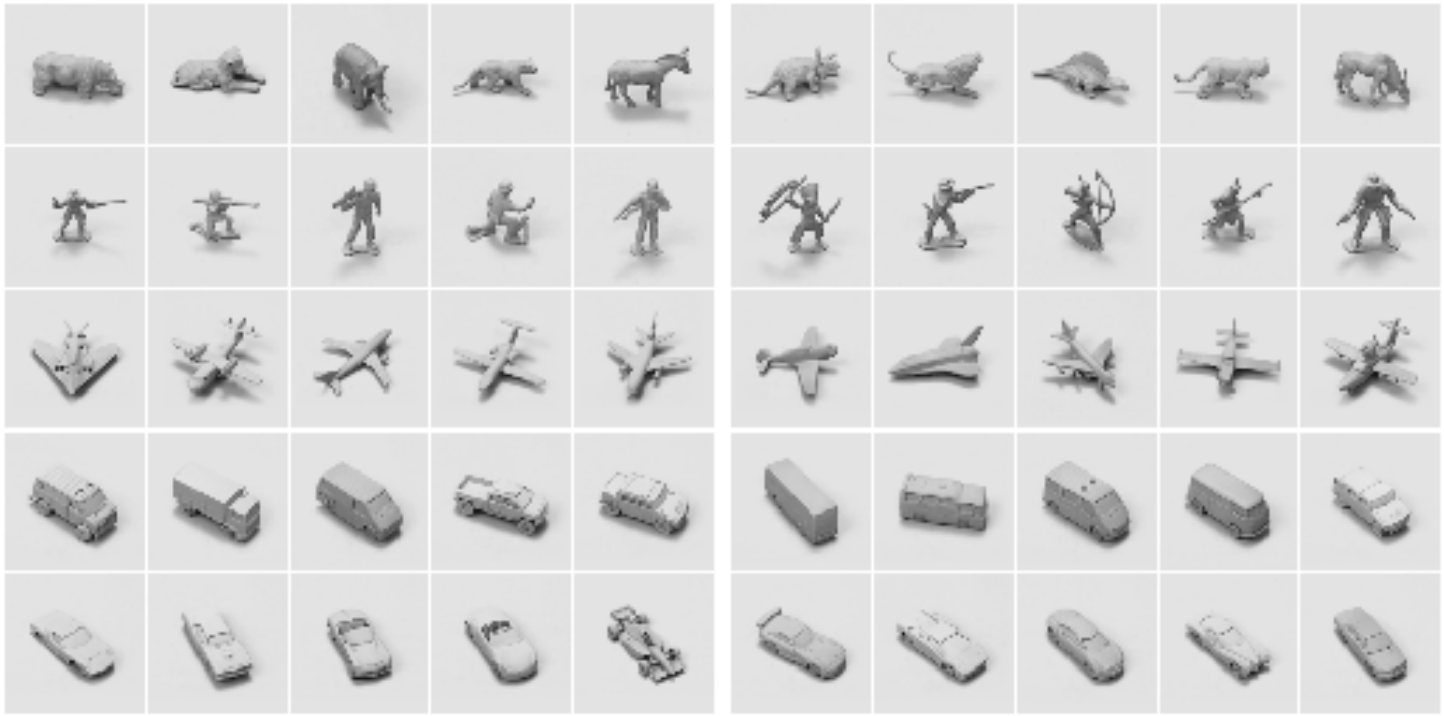
Limits

Numbers worked great, and so did faces…

Why did other categories fail?

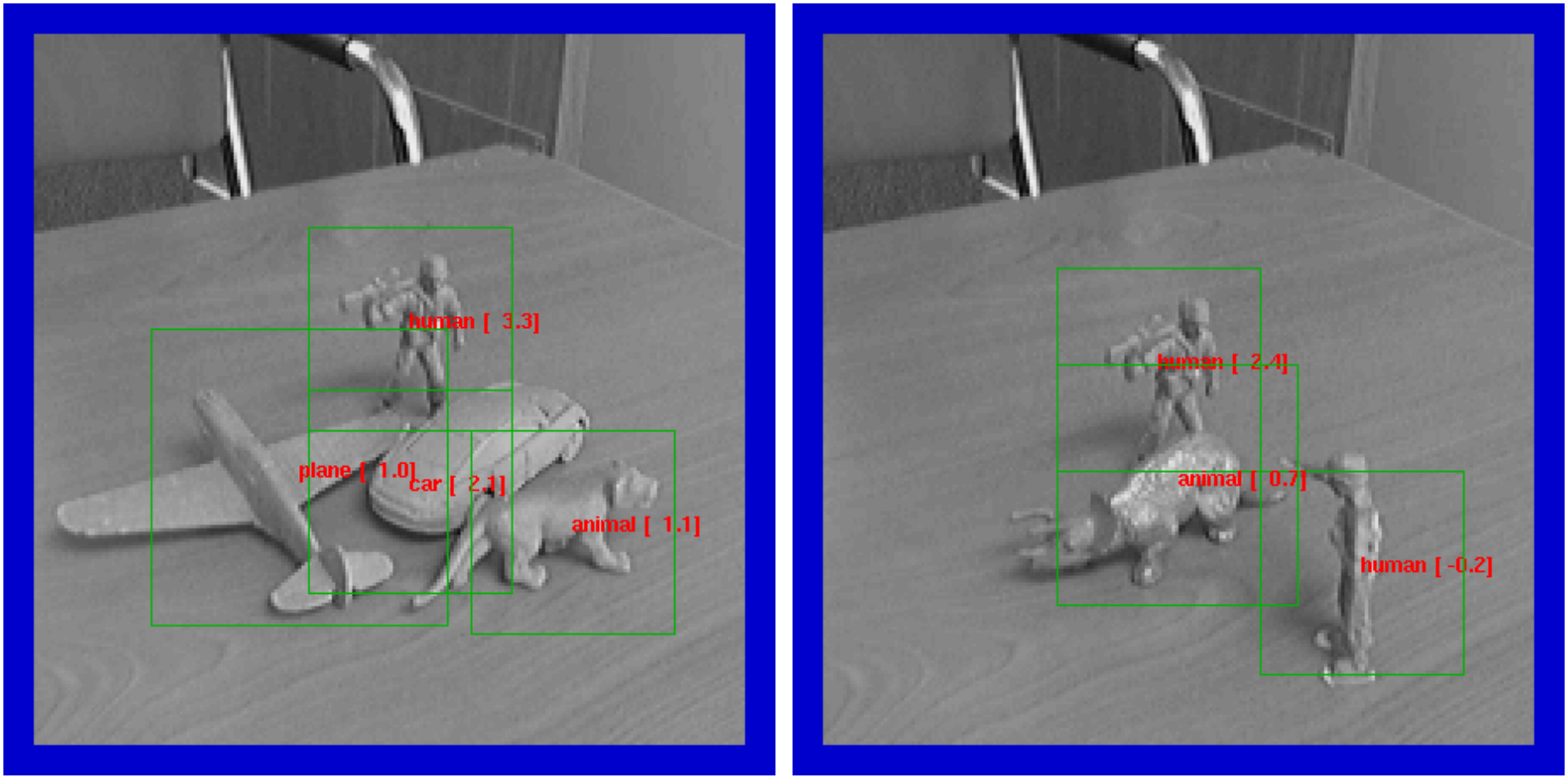
3 main reasons… (2012)

2003 NORB

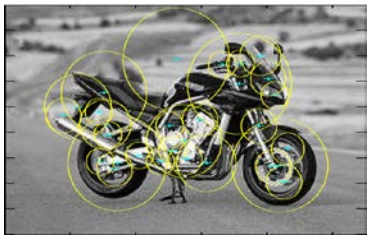


Train

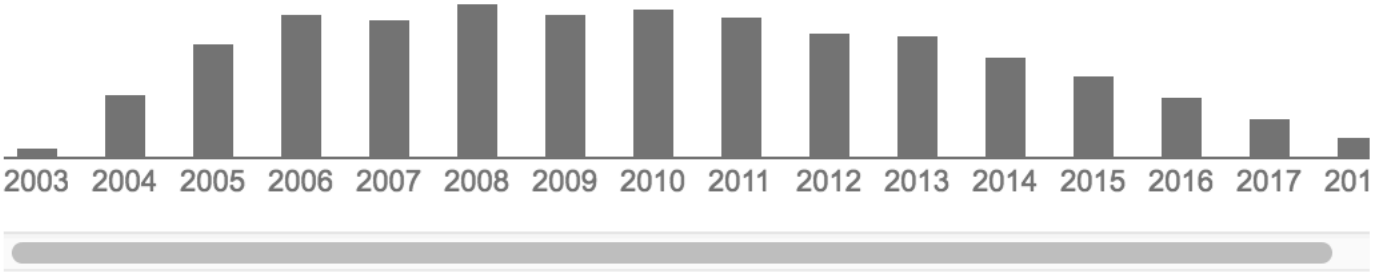
Test



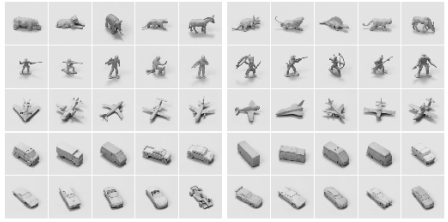
Learning Methods for Generic Object Recognition with Invariance to Pose and Lighting,
LeCun, Huang, Buttou, CVPR 2004



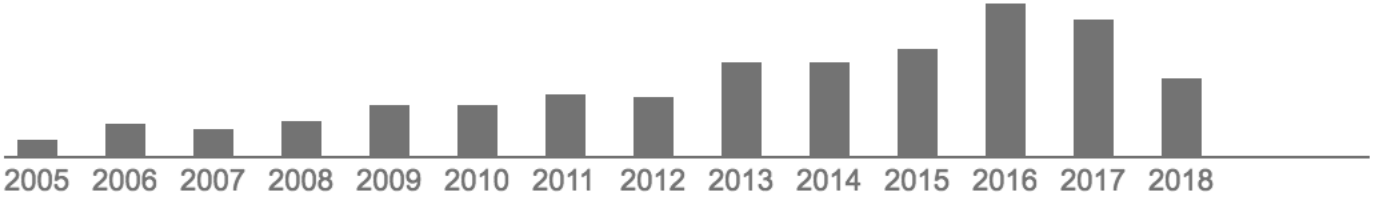
Total citations Cited by 2646



Scholar articles Object class recognition by unsupervised scale-invariant learning
R Fergus, P Perona, A Zisserman - Computer Vision and Pattern Recognition, 2003 …, 2003

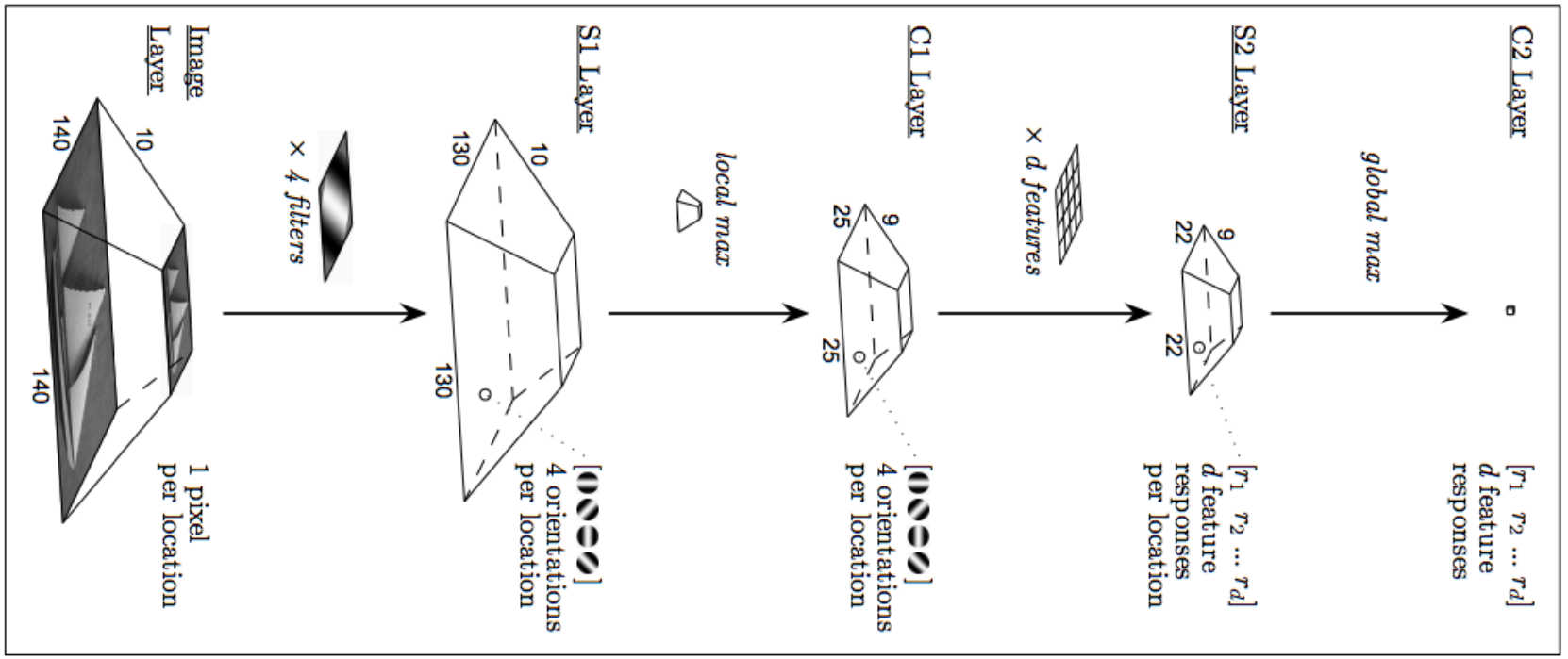


Total citations Cited by 895



Scholar articles Learning methods for generic object recognition with invariance to pose and lighting
Y LeCun, FJ Huang, L Bottou - Computer Vision and Pattern Recognition, 2004.

2006 More clues…



Multiclass Object Recognition with Sparse, Localized Features, Jim Mutch, David Lowe

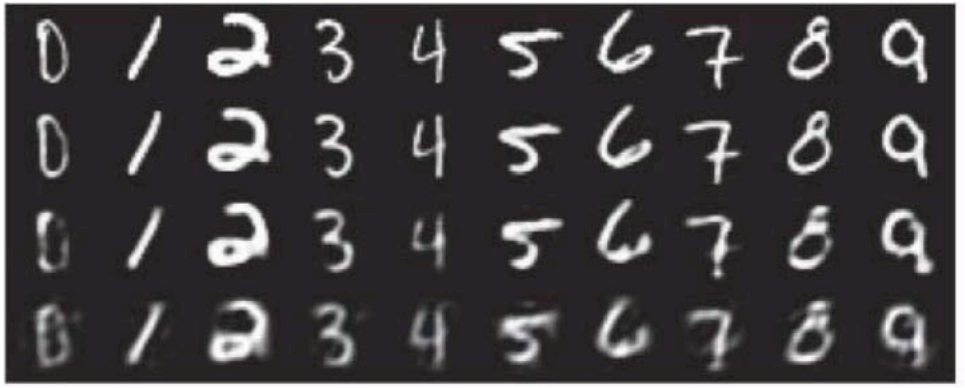
Figure 8. The only 2 errors (1 missed detection, 1 false positive) made in 8 runs on the single-scale UIUC car dataset.

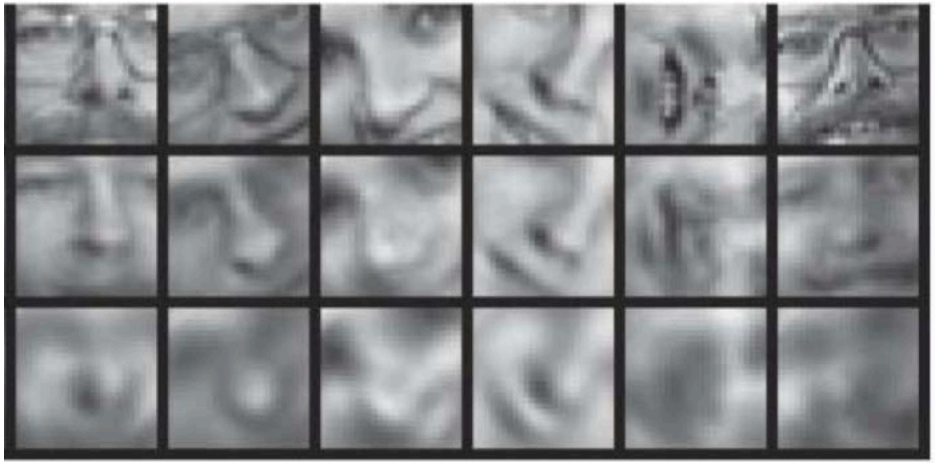
Multiclass Object Recognition with Sparse, Localized Features,
Jim Mutch, David Lowe

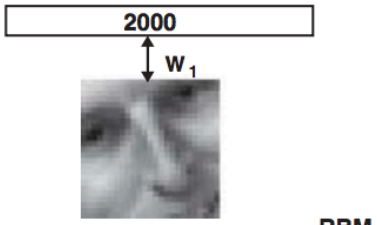
2006

How can we train deeper networks?

Learn one layer at a time…

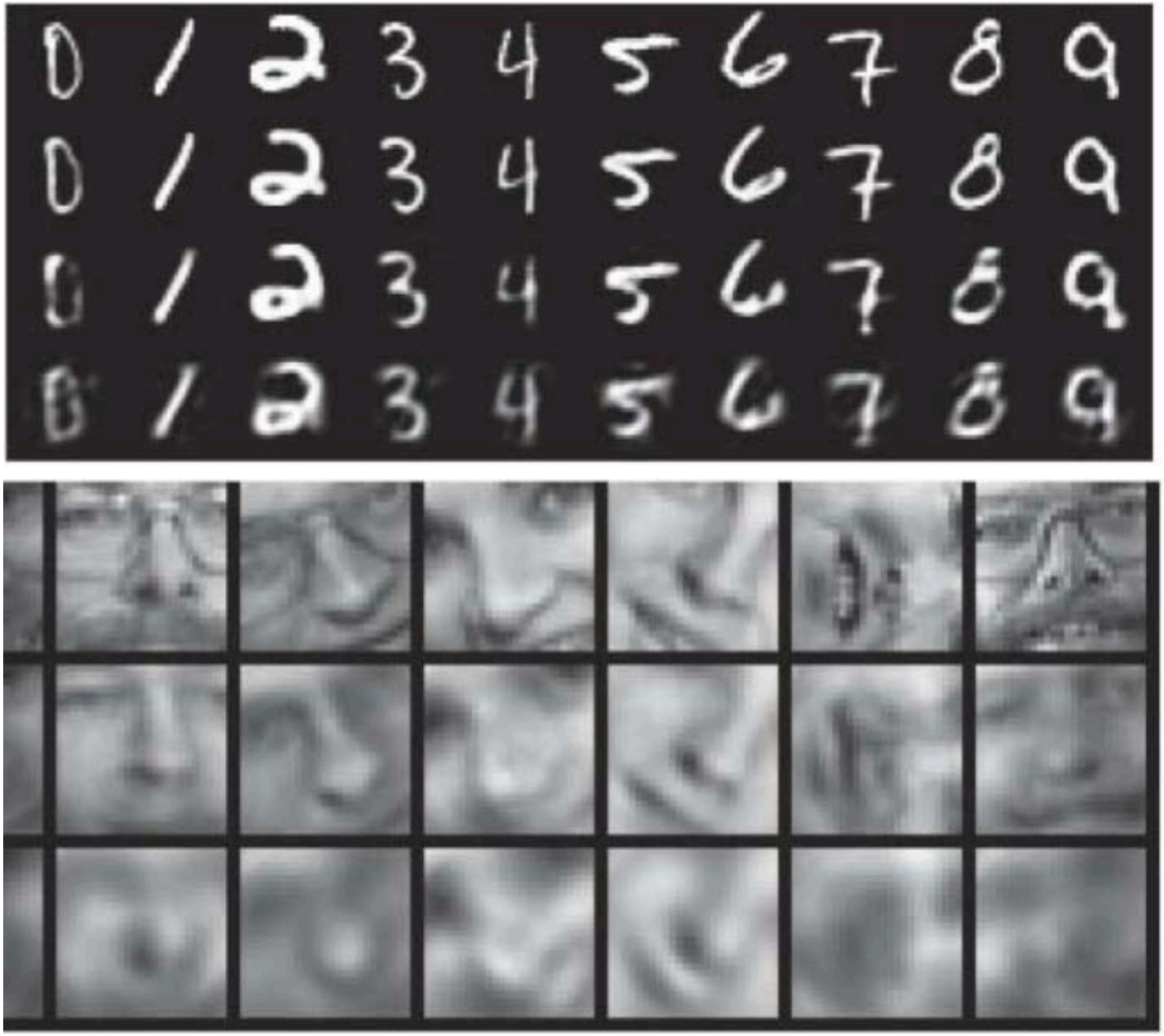




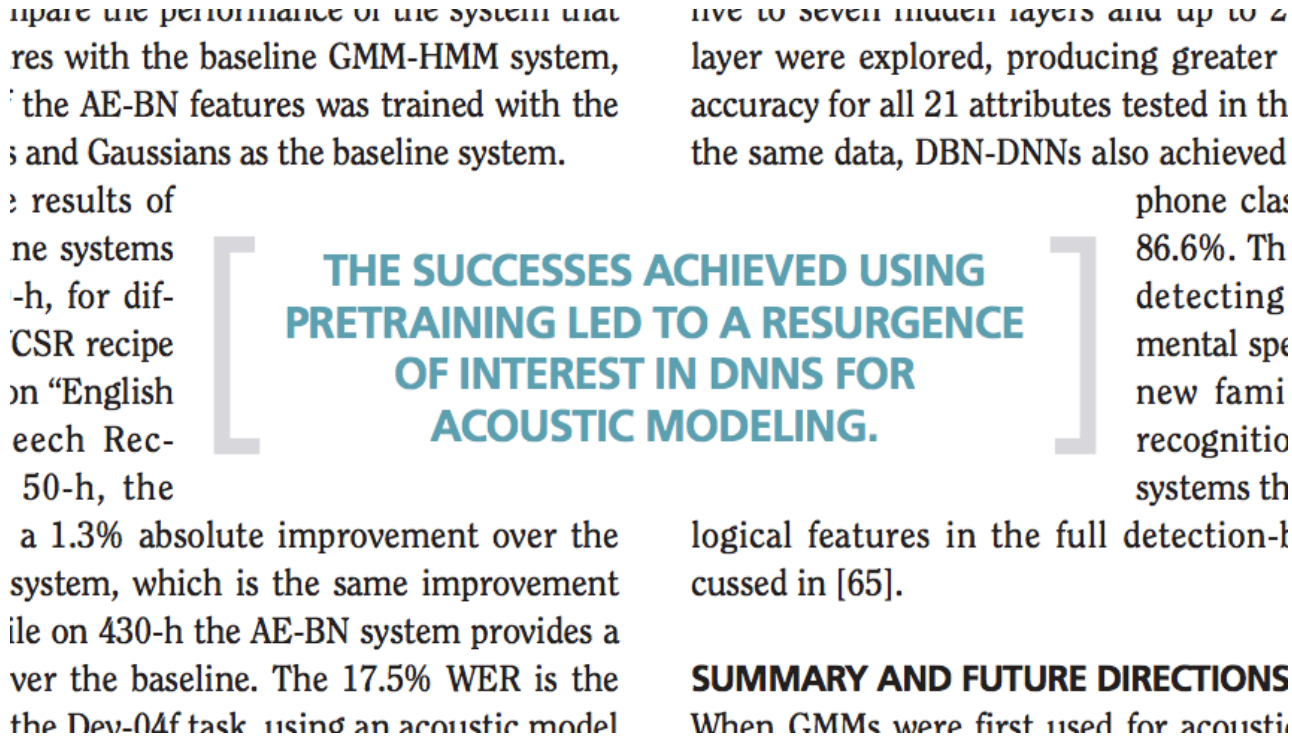


Pretraining

Reducing the Dimensionality of Data with Neural Networks,
G. E. Hinton and R. R. Salakhutdinov, Science 2006



2009-2012 Speech recognition



ıpare the performance of the system that
res with the baseline GMM-HMM system,
the AE-BN features was trained with the
s and Gaussians as the baseline system.
e results of
ne systems
-h, for dif-
CSR recipe
on "English
eech Rec-
50-h, the
a 1.3% absolute improvement over the
system, which is the same improvement
ile on 430-h the AE-BN system provides a
ver the baseline. The 17.5% WER is the
the Dev-04f task, using an acoustic model

five to seven hidden layers and up to 2
layer were explored, producing greater
accuracy for all 21 attributes tested in th
the same data, DBN-DNNs also achieved
phone clas
86.6%. Th
detecting
mental spe
new fami
recognitio
systems th
logical features in the full detection-b
cussed in [65].

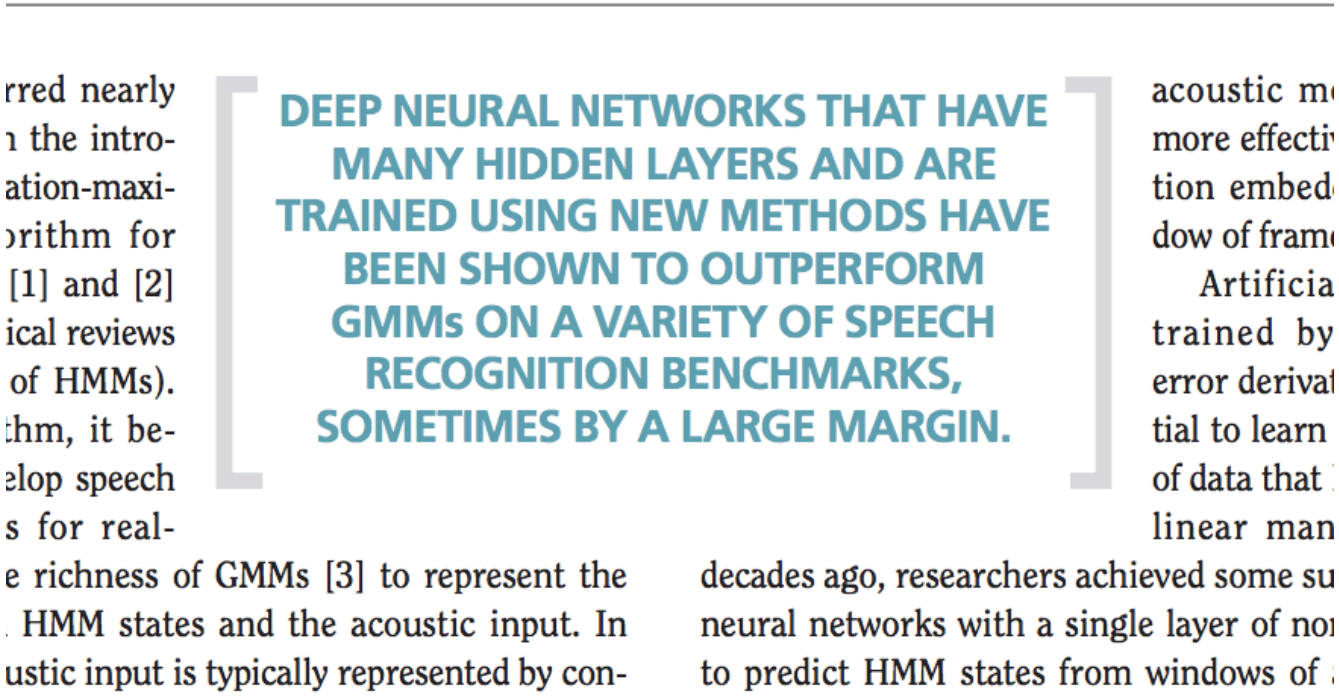
SUMMARY AND FUTURE DIRECTIONS

When GMMs were first used for acoustic

THE SUCCESSES ACHIEVED USING PRETRAINING LED TO A RESURGENCE OF INTEREST IN DNNS FOR ACOUSTIC MODELING.

Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, Hinton et al., 2012

2009-2012 Speech recognition



rred nearly
n the intro-
ation-maxi-
orithm for
[1] and [2]
ical reviews
of HMMs).
thm, it be-
elop speech
s for real-

DEEP NEURAL NETWORKS THAT HAVE MANY HIDDEN LAYERS AND ARE TRAINED USING NEW METHODS HAVE BEEN SHOWN TO OUTPERFORM GMMs ON A VARIETY OF SPEECH RECOGNITION BENCHMARKS, SOMETIMES BY A LARGE MARGIN.

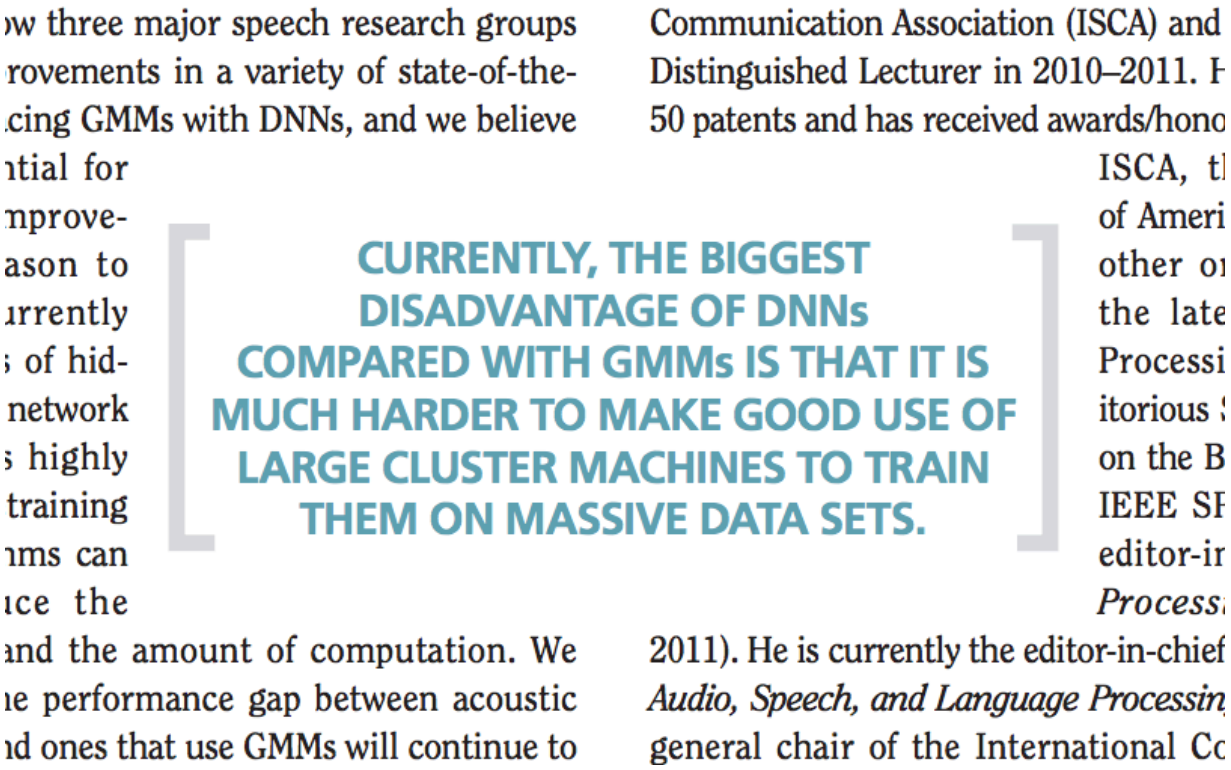
acoustic m
more effecti
tion embed
dow of fram
Artificia
trained by
error derivat
tial to learn
of data that
linear man

e richness of GMMs [3] to represent the
HMM states and the acoustic input. In
ustic input is typically represented by con-

decades ago, researchers achieved some su
neural networks with a single layer of no
to predict HMM states from windows of

Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, Hinton et al., 2012

2009-2012 Speech recognition



ɔw three major speech research groups
rovements in a variety of state-of-the-
.cing GMMs with DNNs, and we believe
ıtial for
ɪmprove-
ason to
ırrently
ɜ of hid-
network
ɜ highly
training
ıms can
ıce the
and the amount of computation. We
ıe performance gap between acoustic
ıd ones that use GMMs will continue to

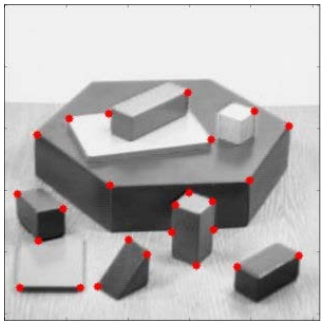
CURRENTLY, THE BIGGEST DISADVANTAGE OF DNNs COMPARED WITH GMMs IS THAT IT IS MUCH HARDER TO MAKE GOOD USE OF LARGE CLUSTER MACHINES TO TRAIN THEM ON MASSIVE DATA SETS.

Communication Association (ISCA) and
Distinguished Lecturer in 2010–2011. H
50 patents and has received awards/hono
ISCA, tl
of Ameri
other o
the late
Processi
itorious
on the B
IEEE SI
editor-ir
Process
2011). He is currently the editor-in-chief
Audio, Speech, and Language Processin
general chair of the International Co

Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, Hinton et al., 2012

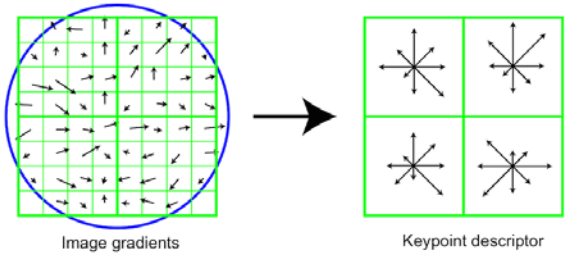
Back to vision…
…25 years of feature designing

Canny Edge Detection
Canny, 1986

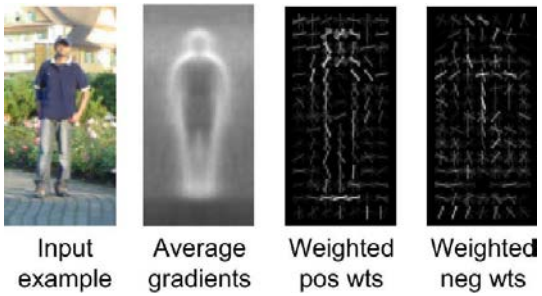


Harris Corner Detection
Harris and Stephens, 1988

Harr Wavelets,
Viola and Jones, CVPR 2001



SIFT,
Lowe, 2004



HOG,
Dalal and Triggs, 2005

2011

How can I convince vision folks that DNNs work?

Detour: Datasets

How to write a paper:

1. Come up with algorithm.

2. Find/create a dataset that works.

How to write a paper:

1. Pick a dataset.

2. Find an algorithm that works.

2004-2006 Caltech 101 and 256







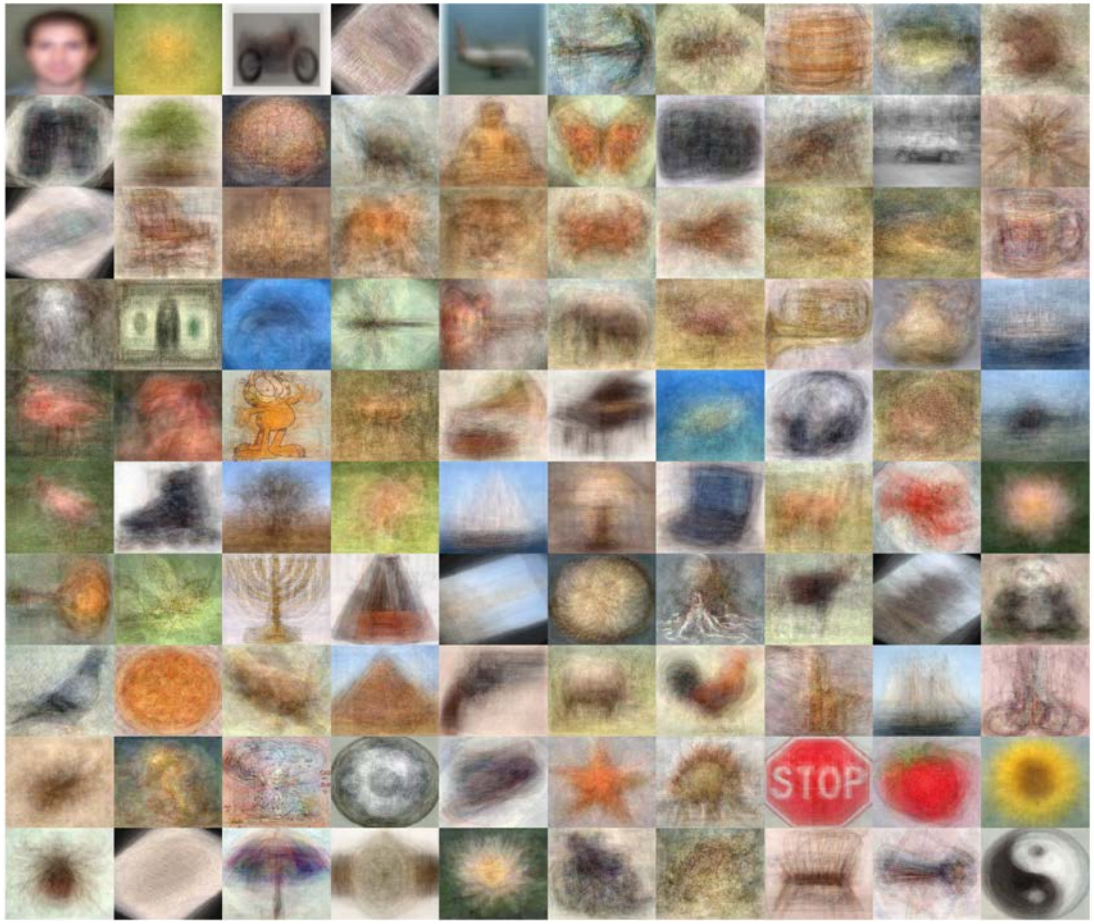







2006

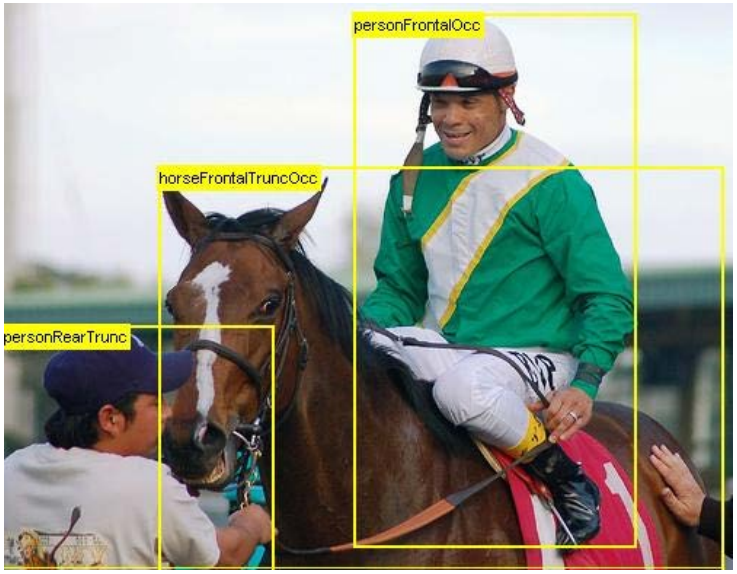
15-30 training images, up to ~70% accuracy.

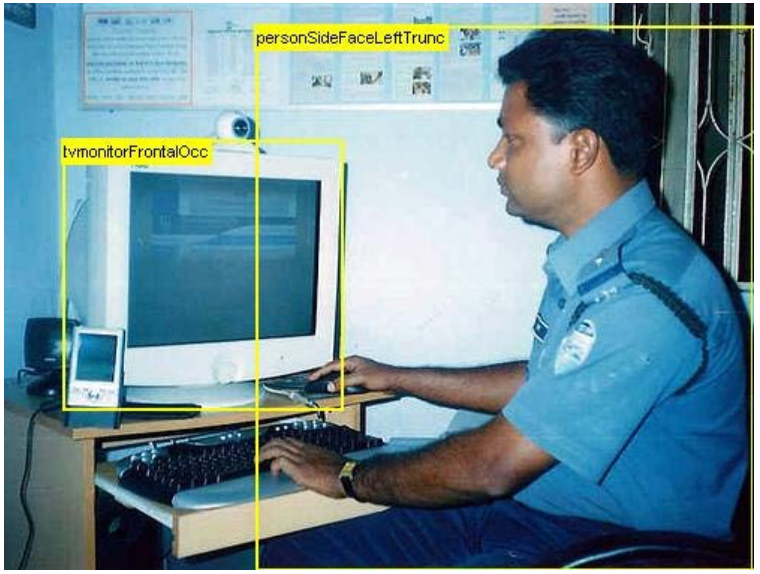


Antonio Torralba

2007 PASCAL VOC

20 classes

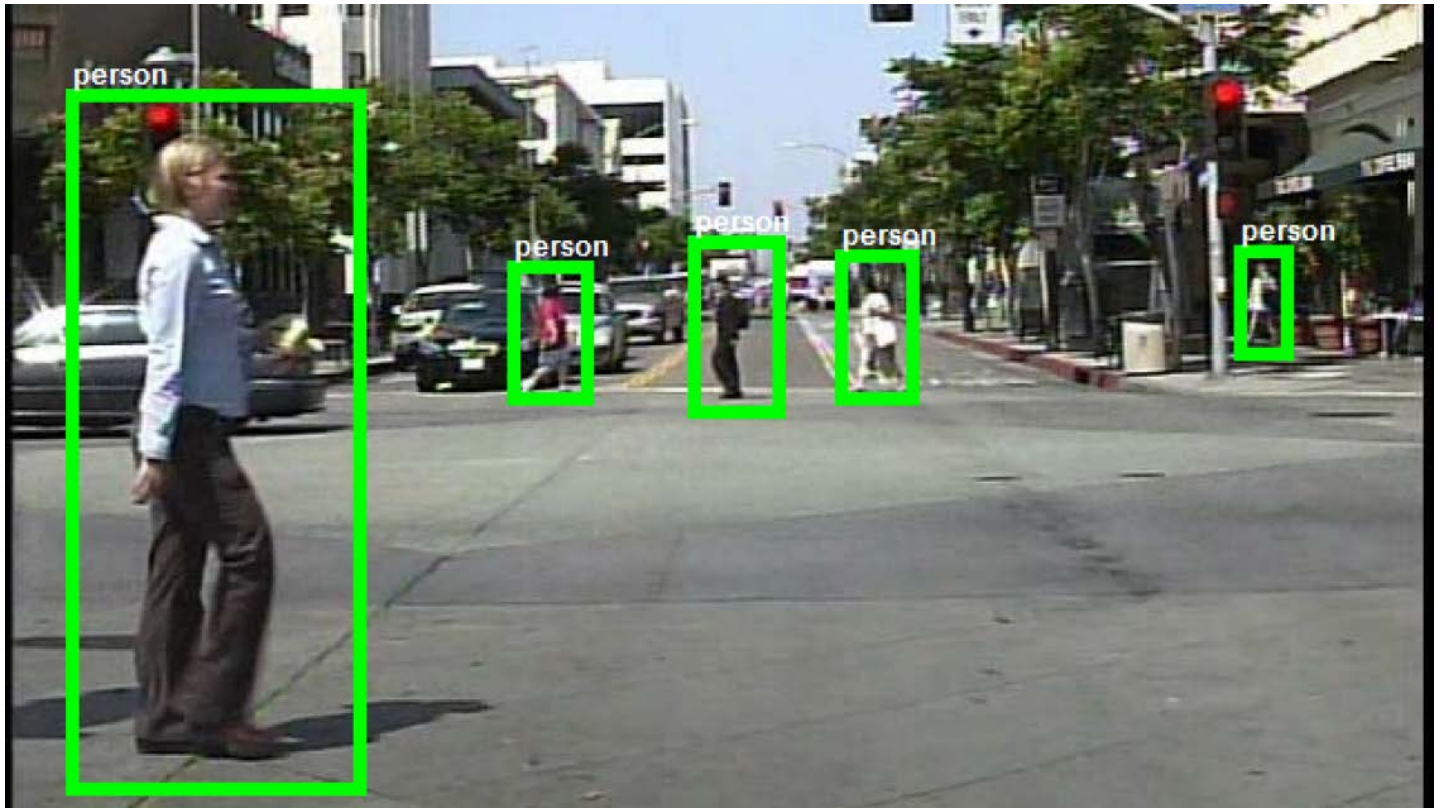




The PASCAL Visual Object Classes (VOC) Challenge, Everingham, Van Gool, Williams, Winn and Zisserman, *IJCV*, 2010

2009 Caltech Pedestrian

1 class, lots of instances.



Pedestrian Detection: An Evaluation of the State of the Art, Dollár, Wojek, Schiele and Perona, *PAMI*, 2012

2009 ImageNet

22K categories, 14M images

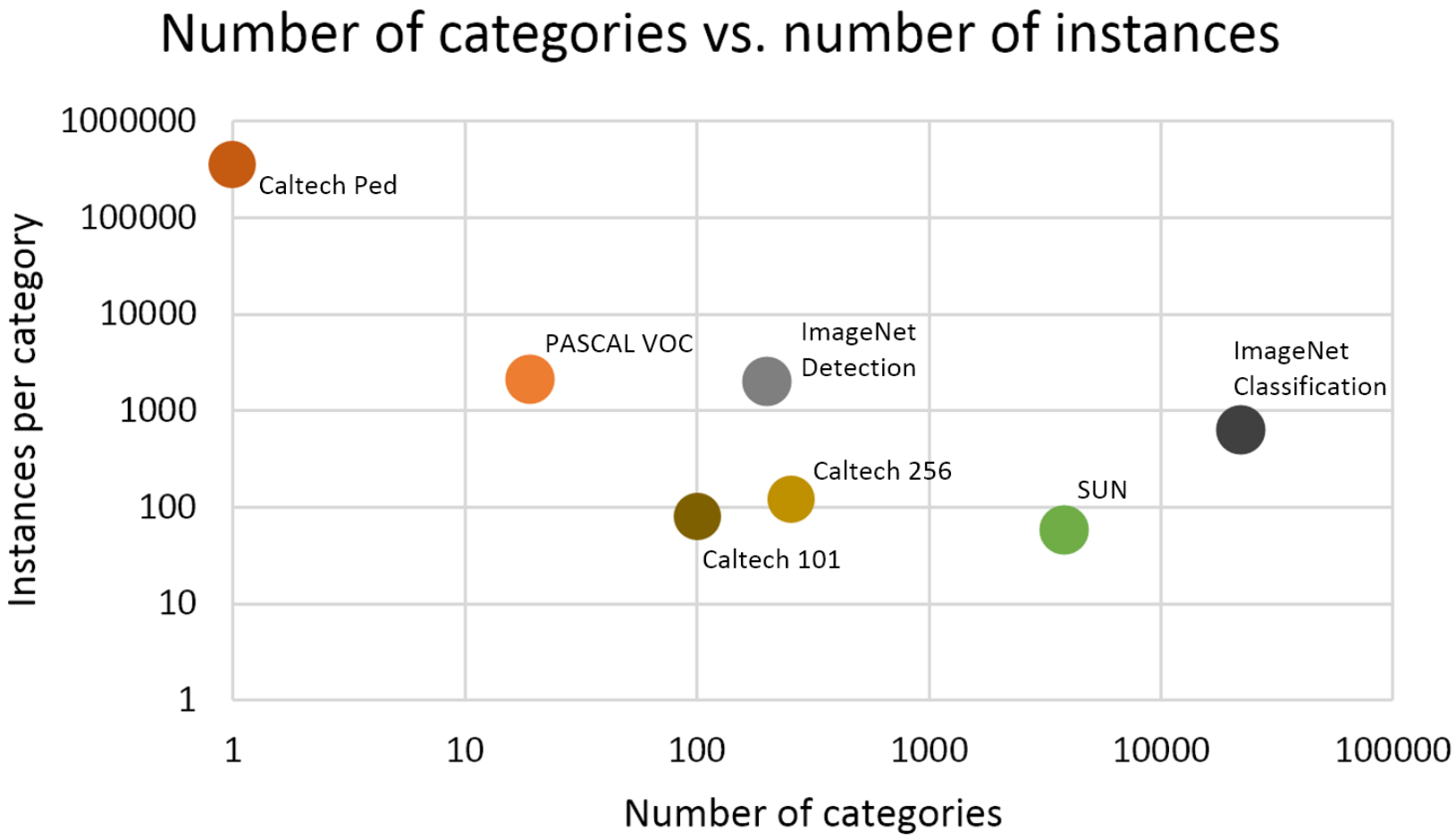
ImageNet: A Large-Scale Hierarchical Image Database,
Deng, Dong, Socher, Li, Li and Fei-Fei, *CVPR,* 2009

2010 SUN

908 scene categories

Beer garden

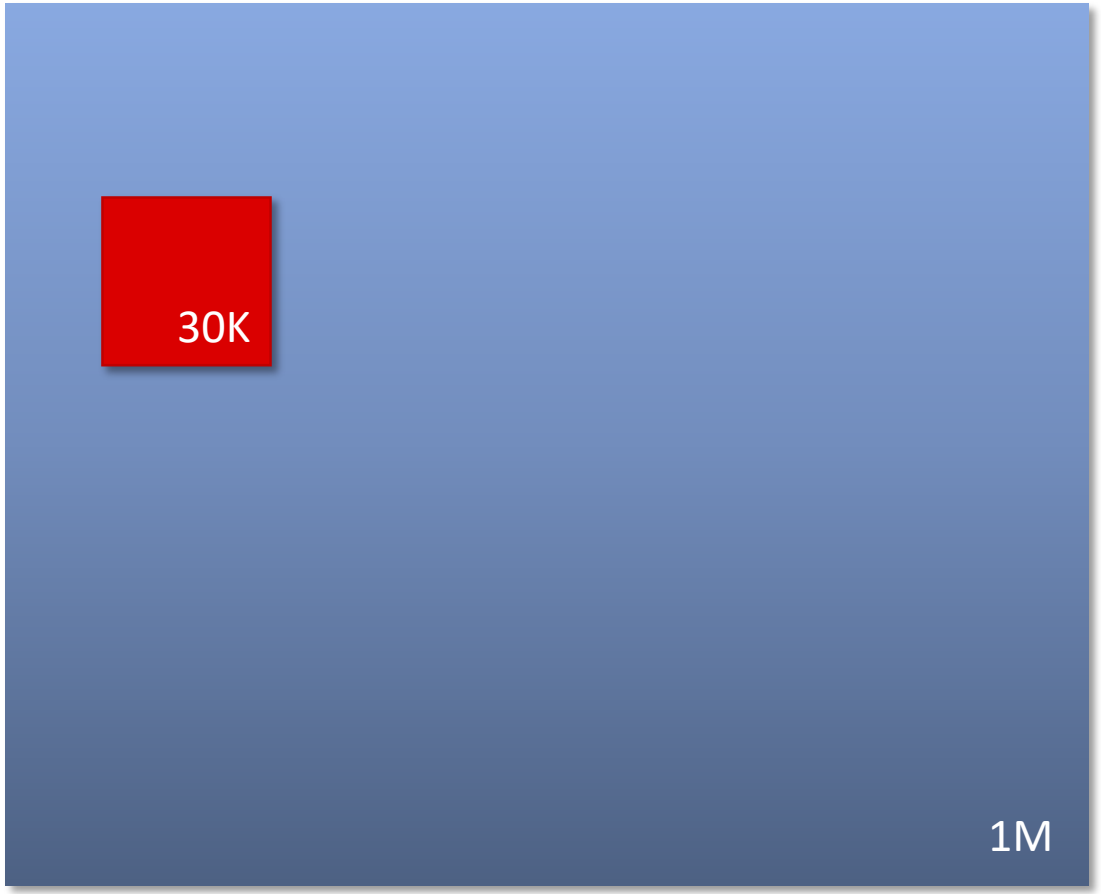
SUN Database: Large-scale Scene Recognition from Abbey to Zoo
Xiao, Hays, Ehinger, Oliva, and Torralba, *CVPR*, 2010.


Number of categories vs. number of instances
Instances per category
1000000
100000
10000
1000
100
10
1
Caltech Ped
PASCAL VOC
ImageNet Detection
ImageNet Classification
Caltech 256
SUN
Caltech 101
1
10
100
1000
10000
100000
Number of categories

Images

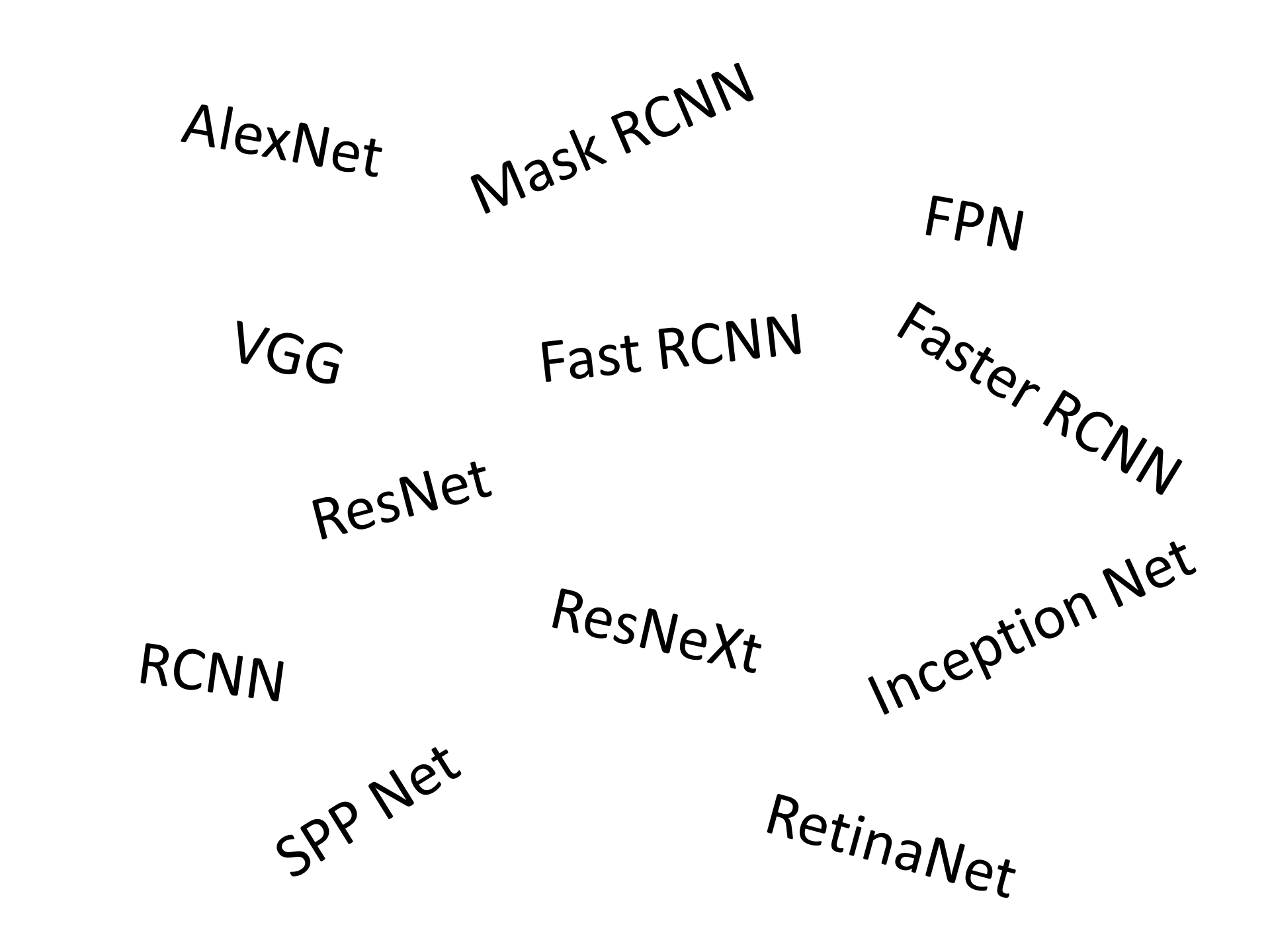
2009

2012



ImageNet

Back to algorithms…


AlexNet
Mask RCNN
FPN
VGG
Fast RCNN
Faster RCNN
ResNet
ResNeXt
Inception Net
RCNN
SPP Net
RetinaNet

Imagenet Classification with Deep Convolutional Neural Networks,
Krizhevsky, Sutskever, and Hinton, *NIPS* 2012

Very deep convolutional networks for large-scale image recognition,
Karen Simonyan, Andrew Zisserman, *ICLR* 2015

Going deeper with convolutions,
Szegedy, Liu, Jia, Sermanet, Reed, Anguelov, Erhan, Vanhoucke, Rabinovich, *CVPR* 2015

Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition
Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, *ECCV* 2014

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

Fast R-CNN
Ross Girshick, *ICCV* 2015

Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks
S Ren, K He, R Girshick, J Sun, *NIPS* 2015

Deep Residual Learning for Image Recognition
Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, *CVPR* 2016

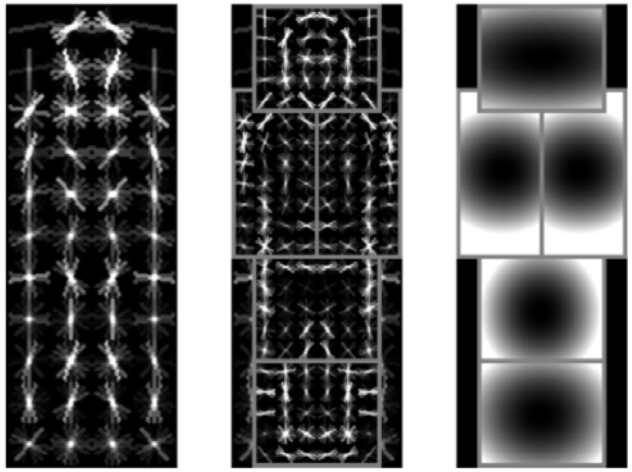
Aggregated residual transformations for deep neural networks
S Xie, R Girshick, P Dollár, Z Tu, K He, *CVPR* 2017

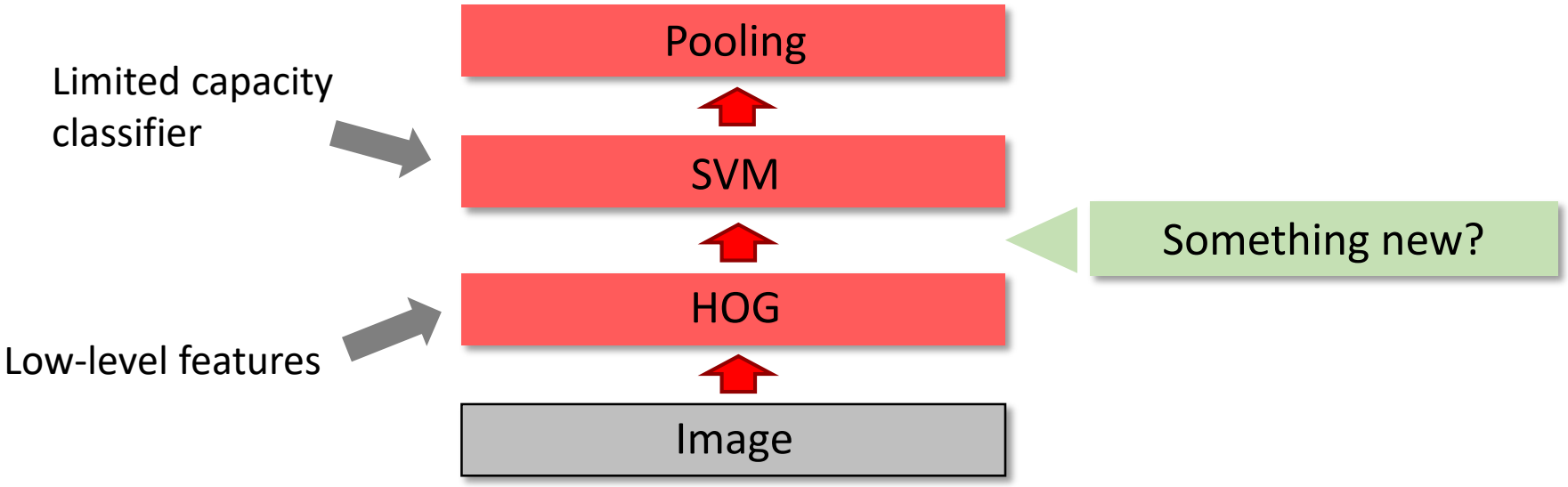
Feature Pyramid Networks for Object Detection,
T. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, S. Belongie, *CVPR* 2017

Mask R-CNN,
Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshick, *ICCV* 2017

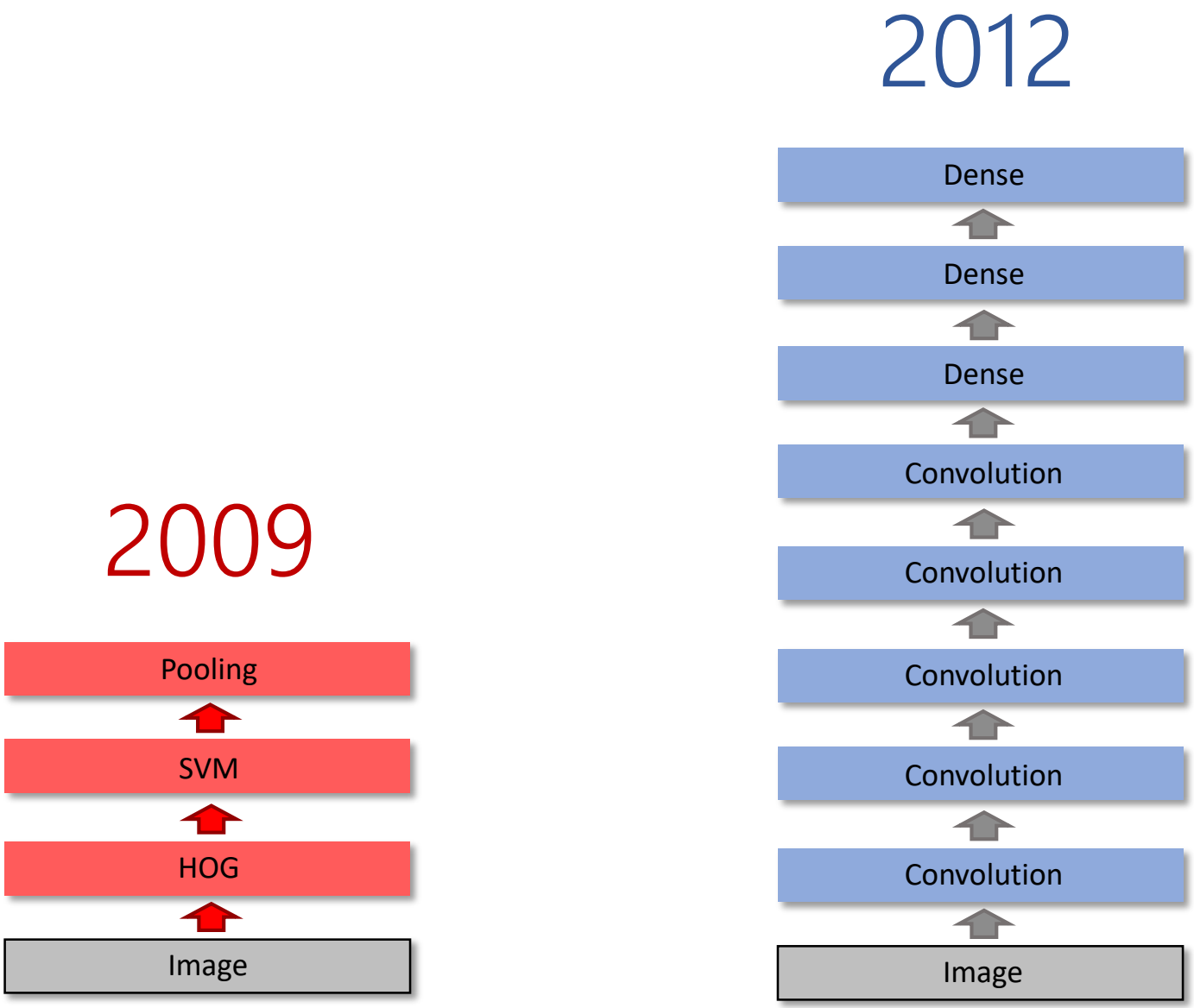
Focal Loss for Dense Object Detection,
Tsung-Yi Lin Priya Goyal Ross Girshick Kaiming He, Piotr Dollar, *ICCV* 2017

2009 DPM




Pooling
SVM
HOG
Image
Limited capacity classifier
Low-level features
Something new?

Algorithms


2012
Dense
Dense
Dense
Convolution
Convolution
Convolution
Convolution
Convolution
Image
2009
Pooling
SVM
HOG
Image

Data
GPUs
\+ Deep Learning

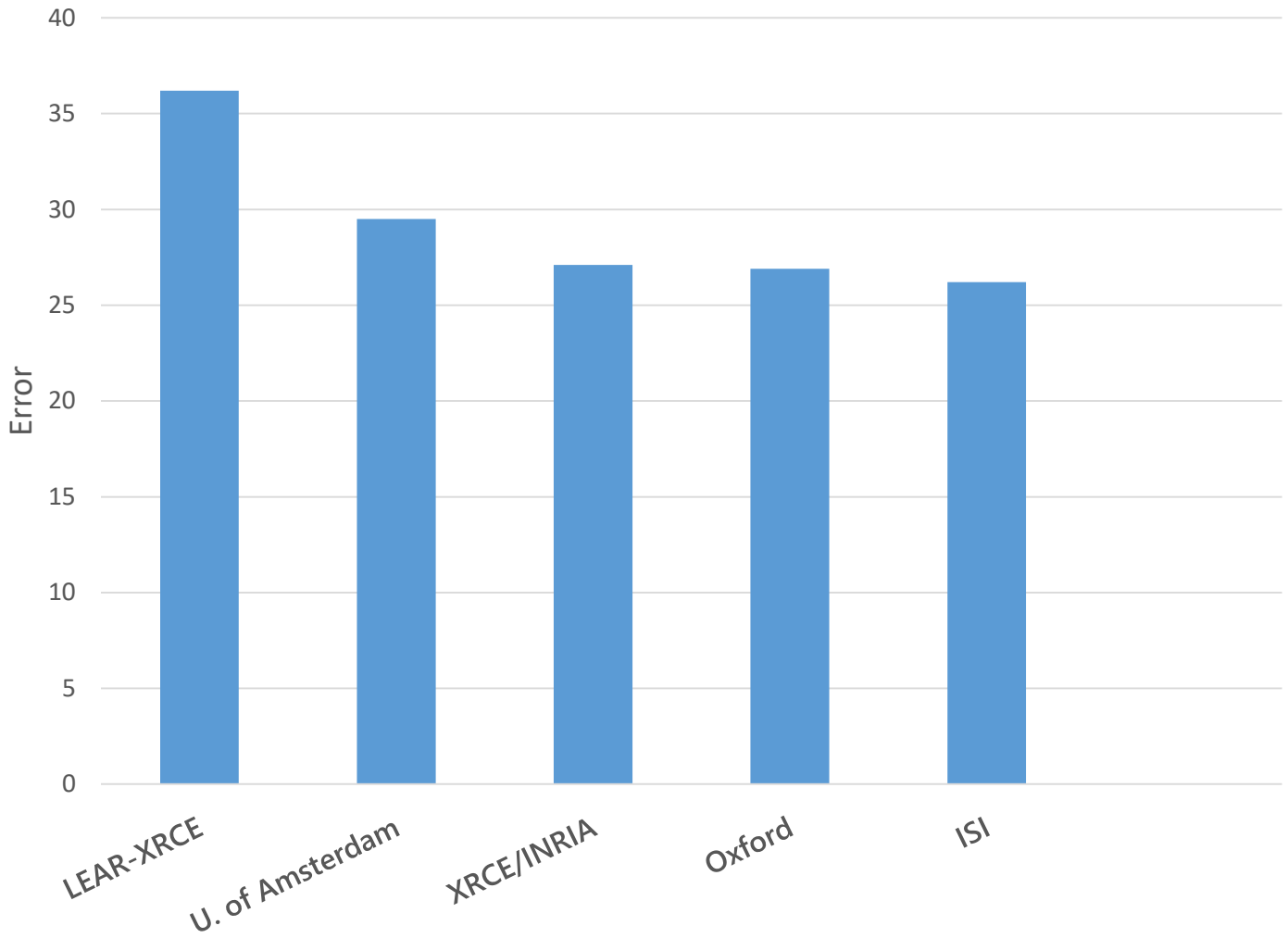
?

2011

How can I convince vision folks that DNNs work?

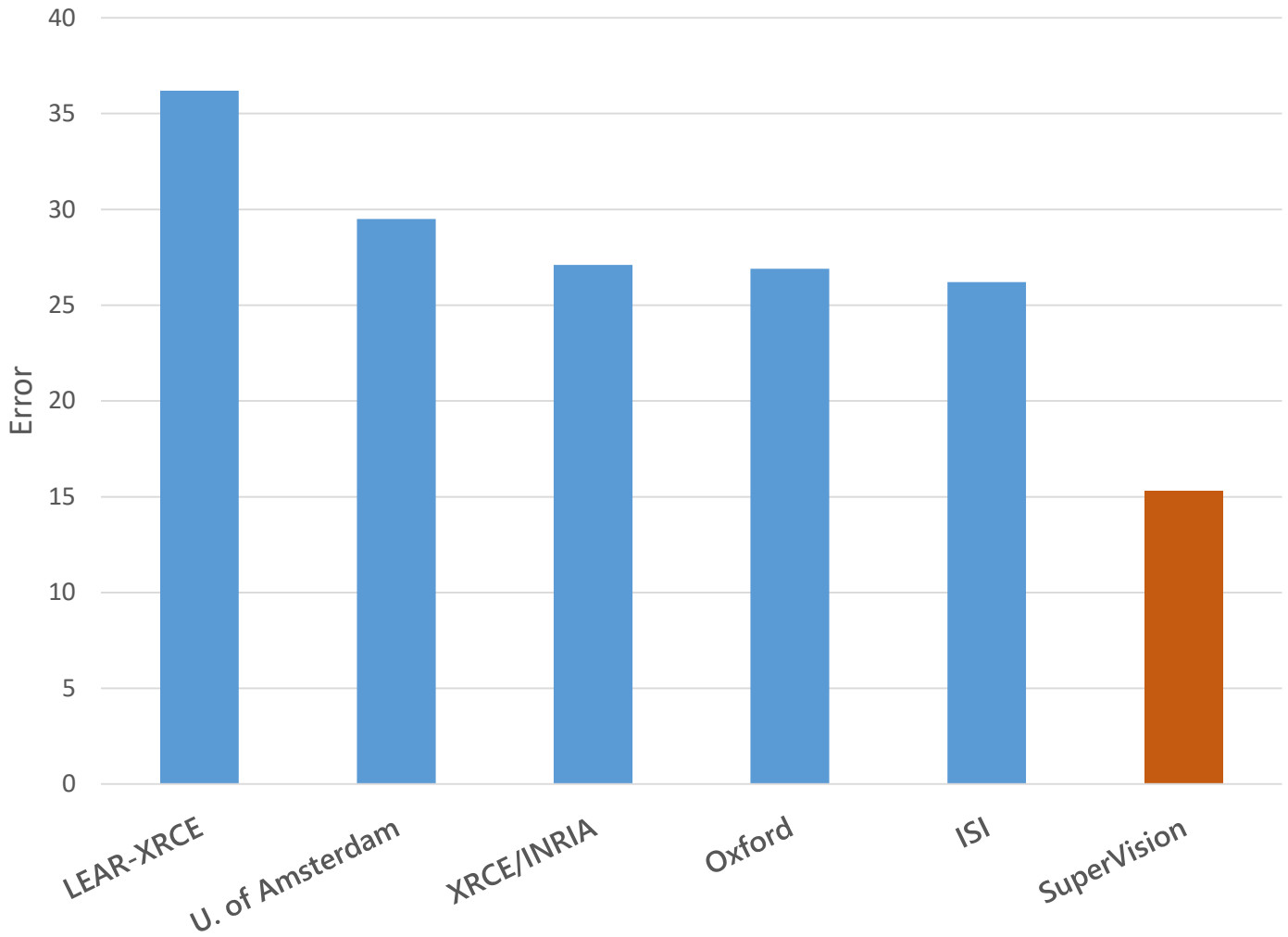
2012 ImageNet 1K

(Fall 2012)



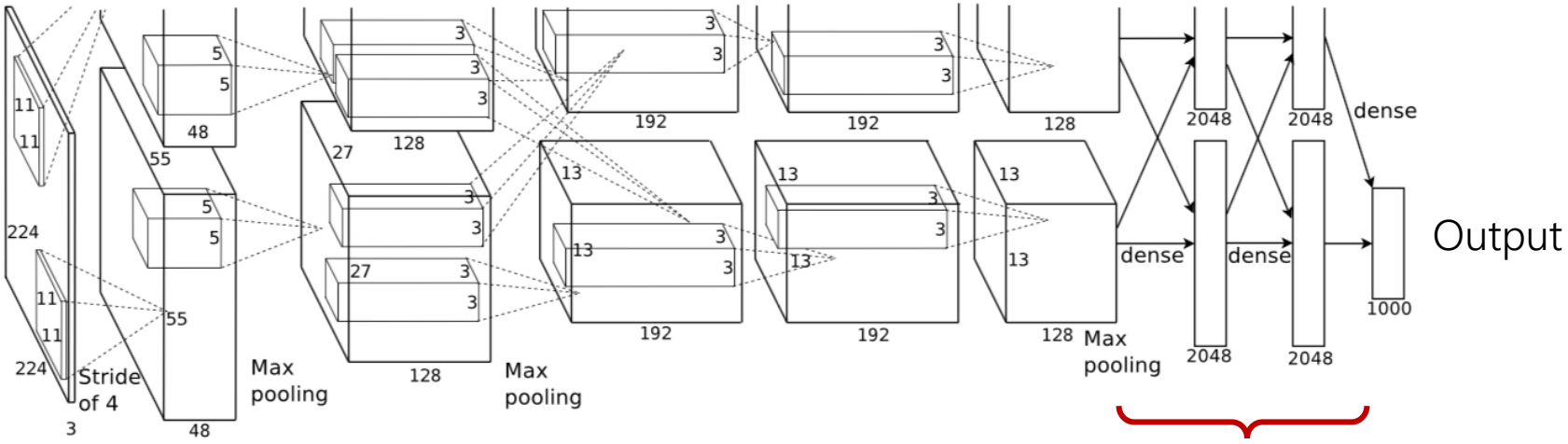
2012 ImageNet 1K

(Fall 2012)



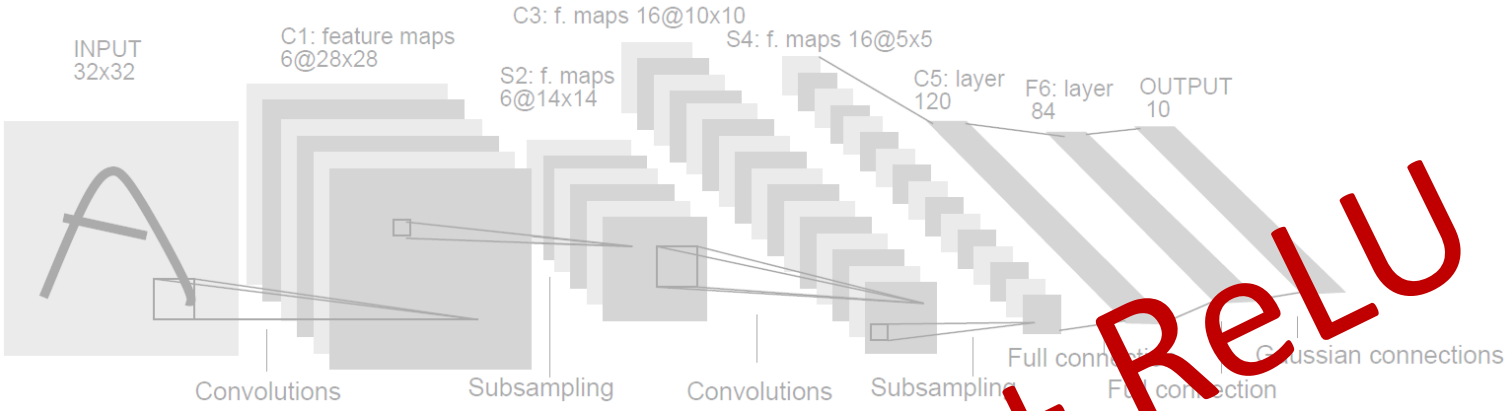
2012 DNNs (deep neural networks)

Image



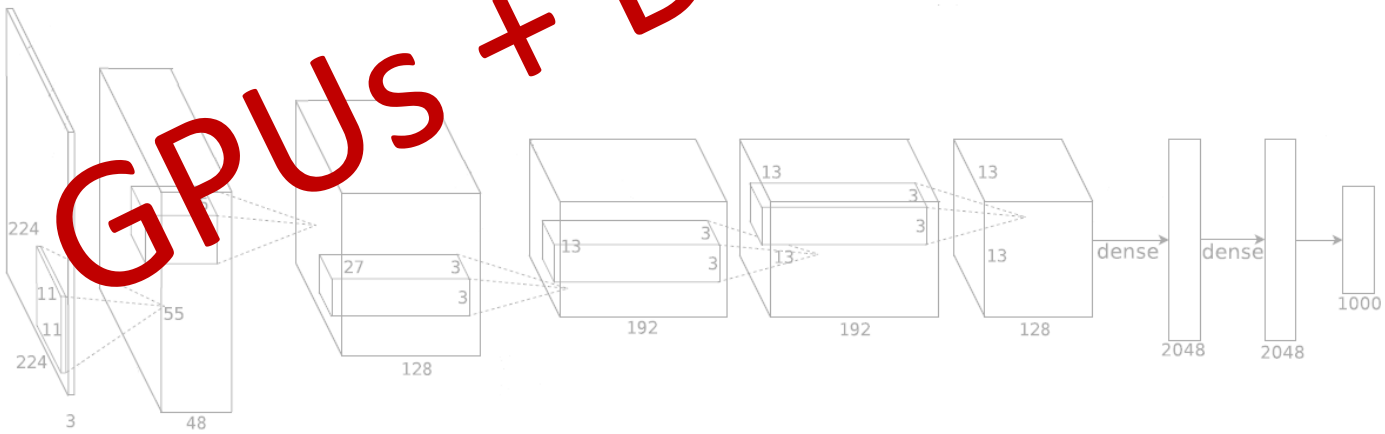
90% parameters

Imagenet Classification with Deep Convolutional Neural Networks, Krizhevsky, Sutskever, and Hinton, NIPS 2012



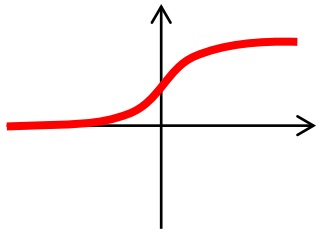
Gradient-Based Learning Applied to Document Recognition,
LeCun, Bottou, Bengio and Haffner, Proc. of the IEEE, **1998**

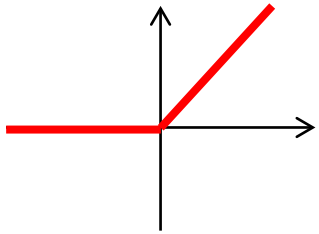
GPUs + Data + ReLU



Imagenet Classification with Deep Convolutional Neural Networks,
Krizhevsky, Sutskever, and Hinton, NIPS **2012**

Why did ReLU make a difference?





step size

derivative of activation function

$$\Delta w_i = \alpha(t_j - y_j) f'(net_j) x_{ij}$$

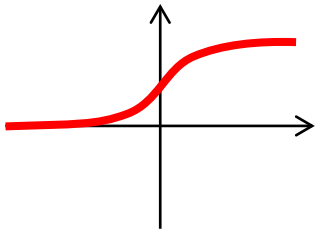
target

output

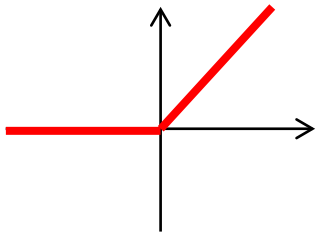
input

$$net_j = \sum_{i=0}^{n} w_i x_{ij}$$

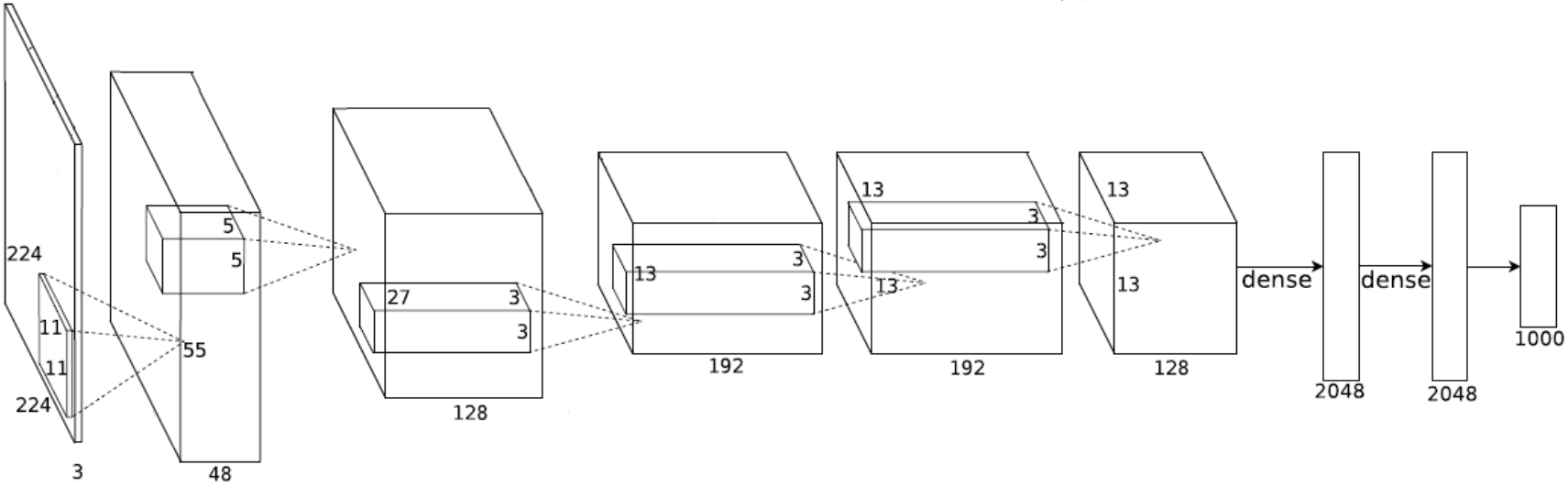
Why did ReLU make a difference?



Sigmoid

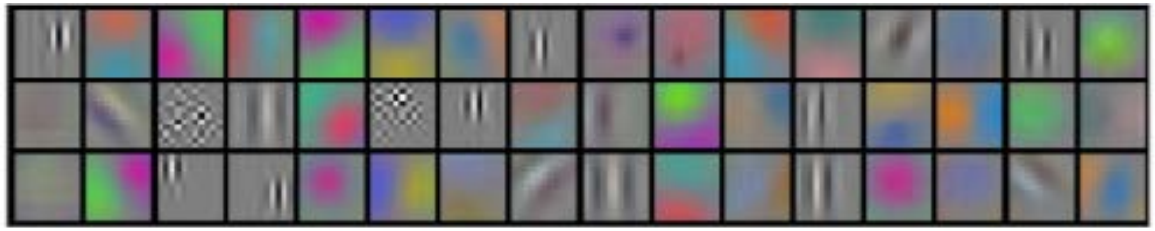


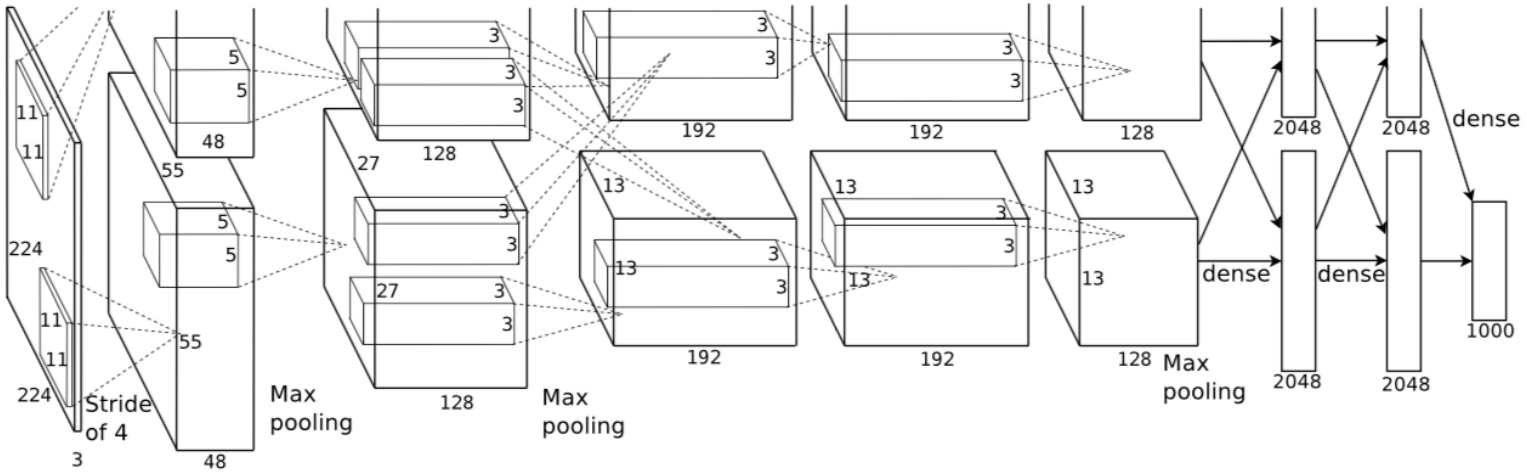
ReLU



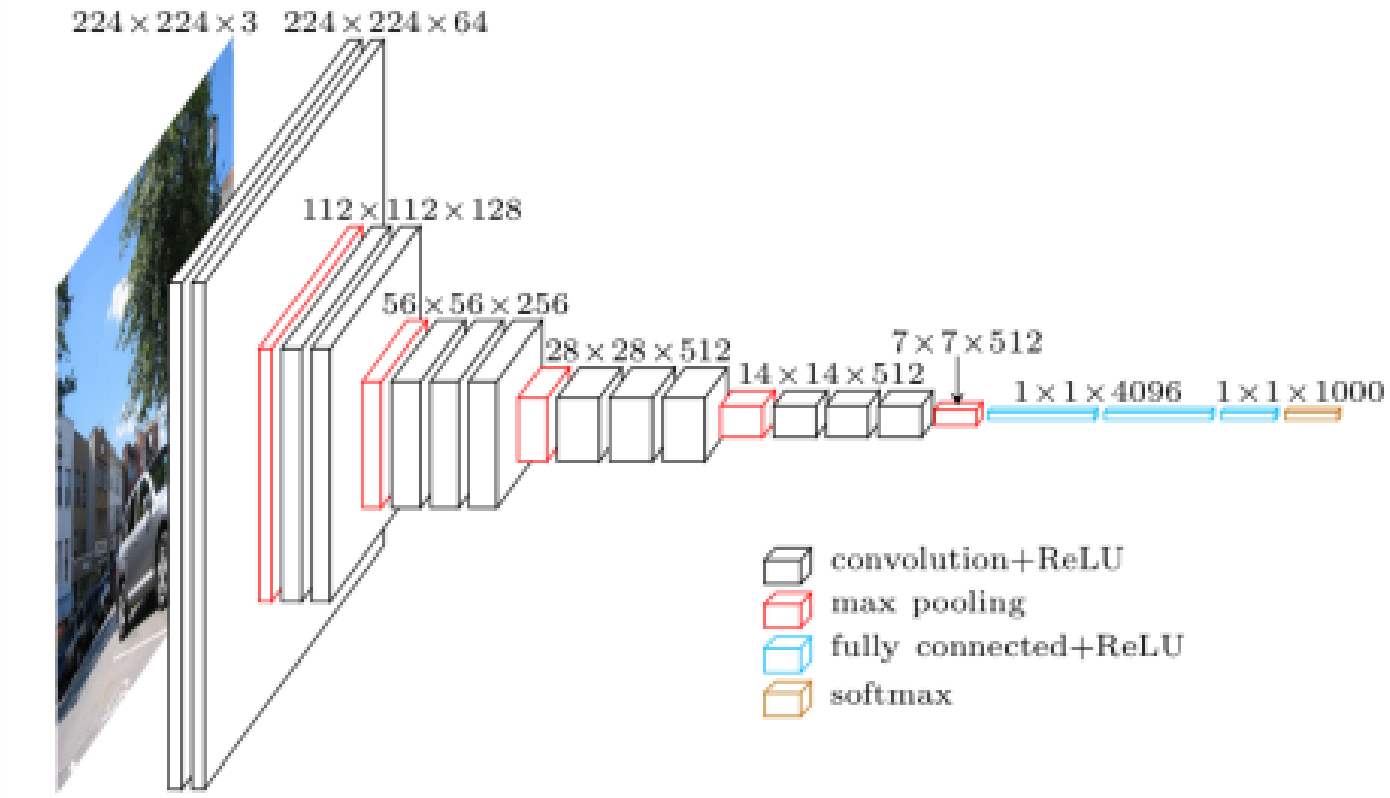
GPU 1

GPU 2




224
224
3
11
11
11
11
Stride
of 4
55
55
48
48
5
5
5
5
Max
pooling
27
27
128
128
3
3
3
3
3
3
Max
pooling
192
192
192
192
128
128
13
13
13
13
13
13
13
13
Max
pooling
2048
2048
2048
2048
dense
dense
dense
1000

2014 VGG Network



Very deep convolutional networks for large-scale image recognition, Karen Simonyan, Andrew Zisserman, ICLR 2015

What made this paper great?

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					

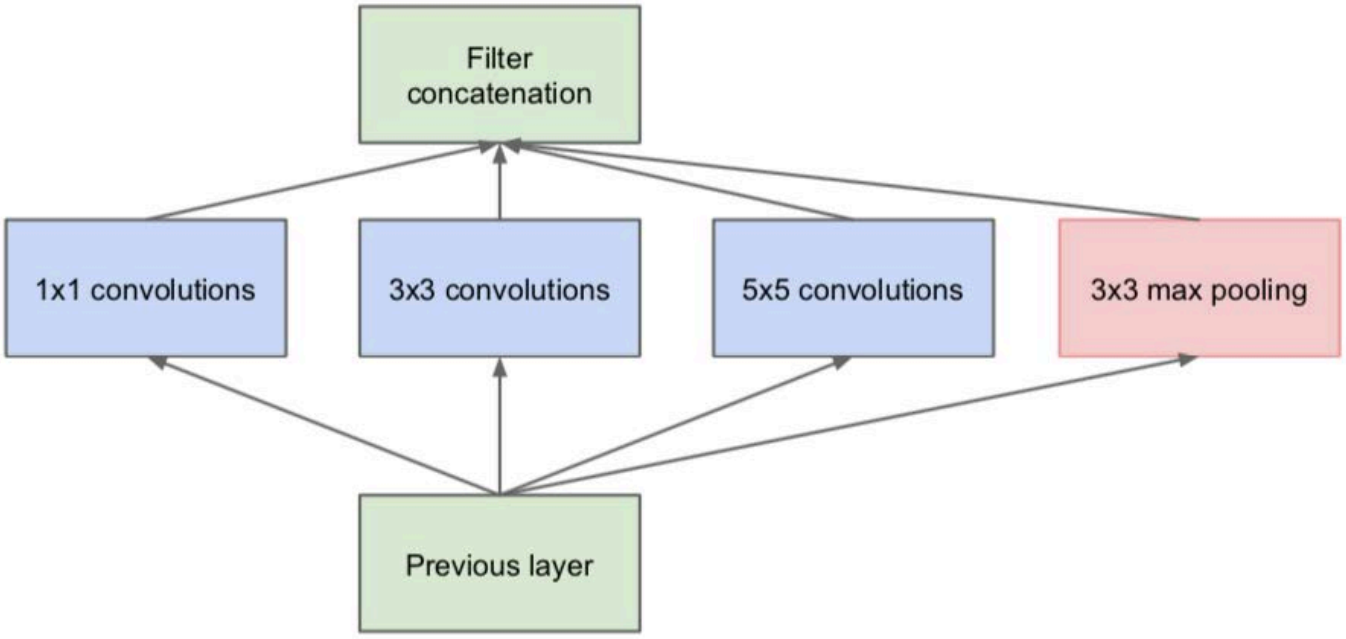
…and it was open sourced

2014 Inception Net

Are there other architecture designs beyond depth and filter size worth exploring?

Going deeper with convolutions,
Szegedy, Liu, Jia, Sermanet, Reed, Anguelov, Erhan, Vanhoucke, Rabinovich, CVPR 2015

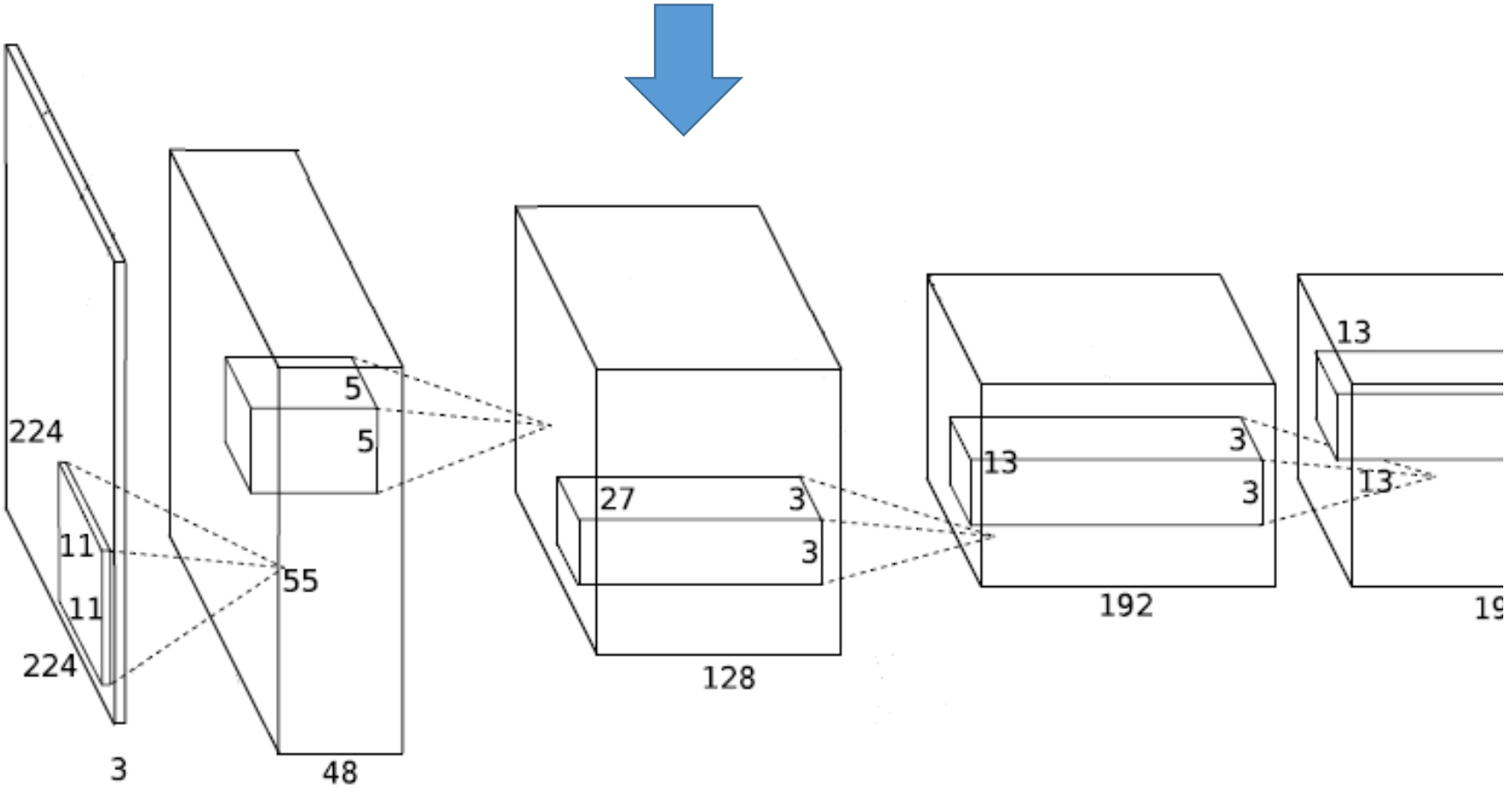
2014 Inception Net



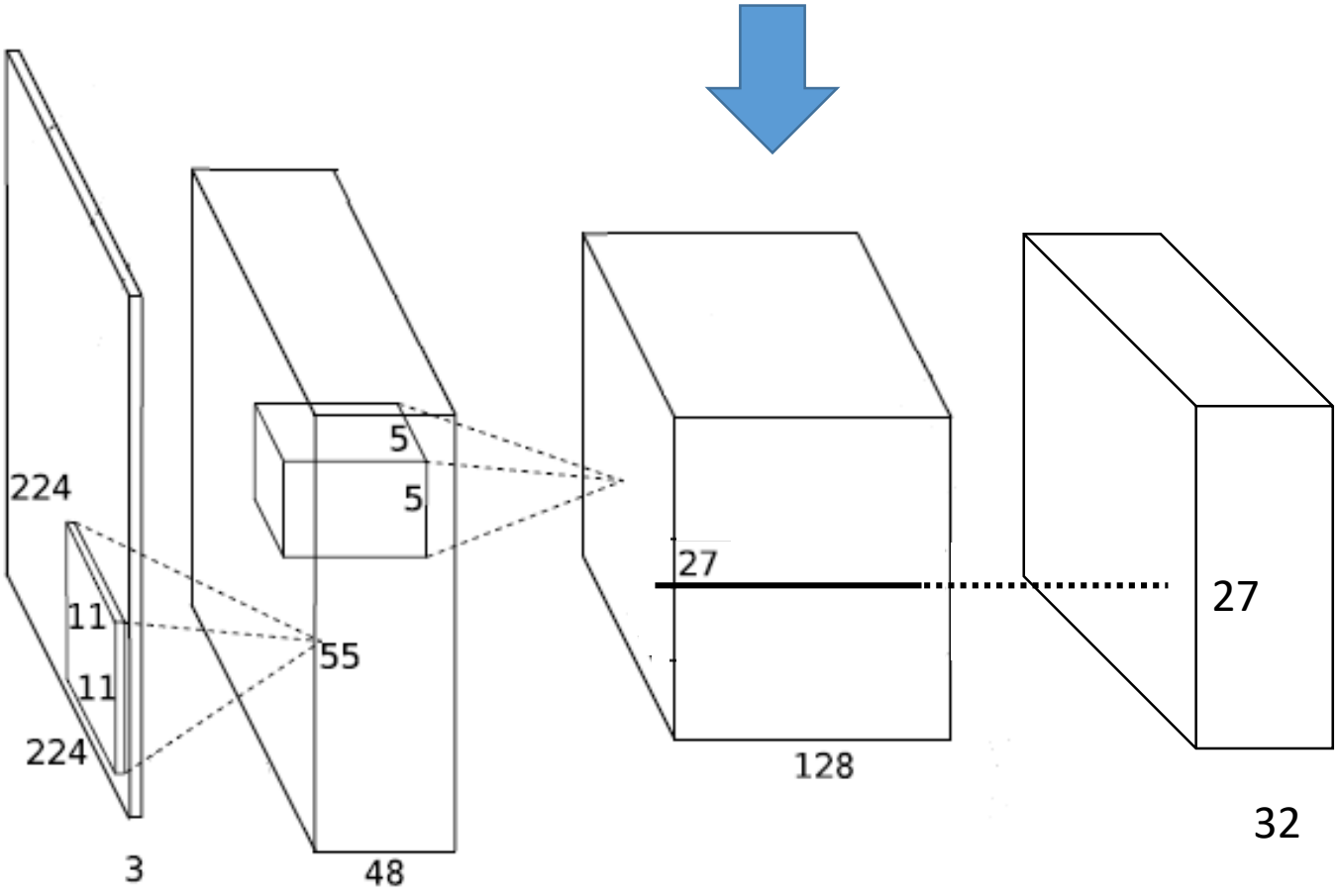
(a) Inception module, naïve version

Going deeper with convolutions,
Szegedy, Liu, Jia, Sermanet, Reed, Anguelov, Erhan, Vanhoucke, Rabinovich, CVPR 2015

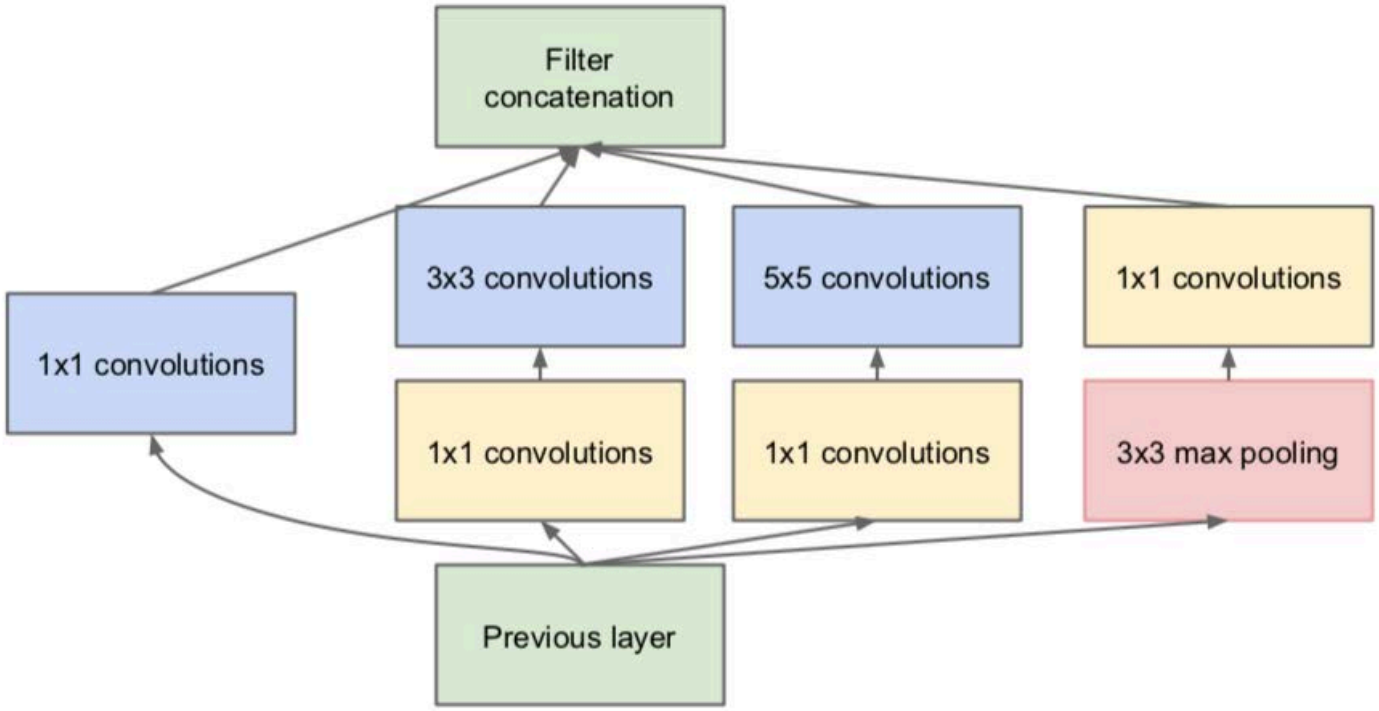
What does it mean to do a 1x1 convolution?



What does it mean to do a 1x1 convolution?



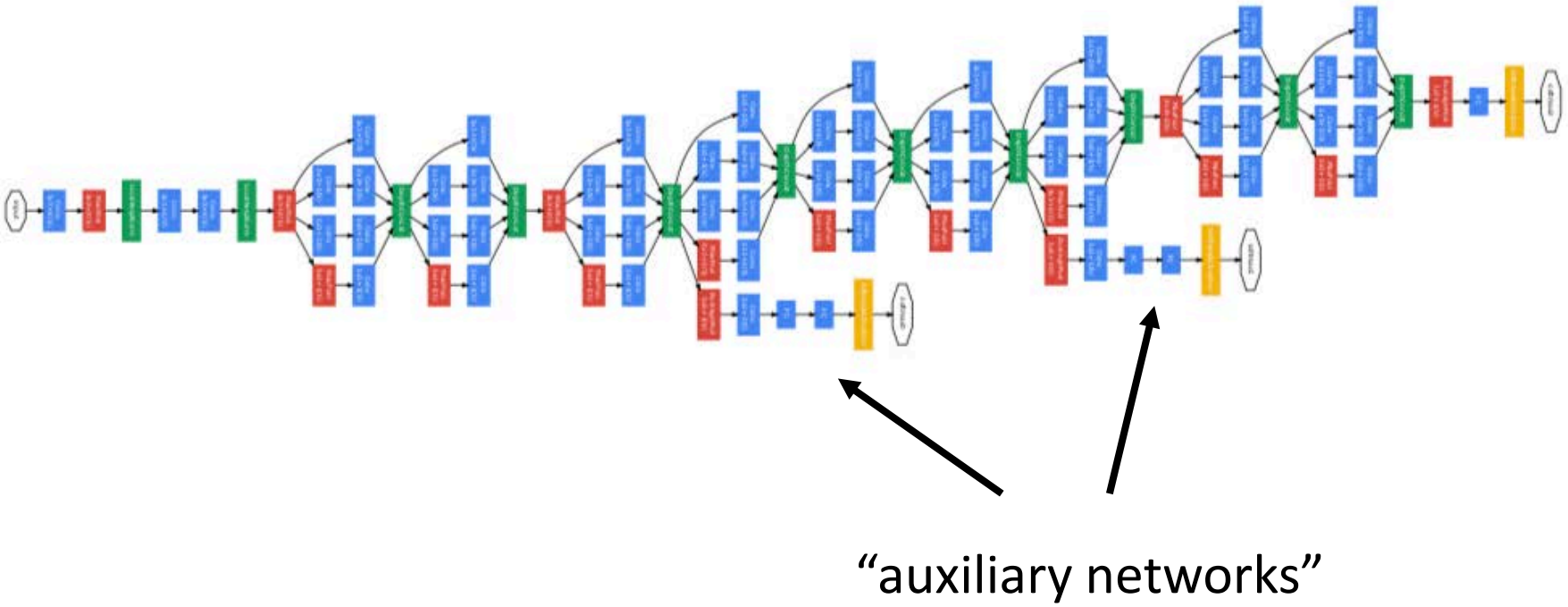
2014 Inception Network



(b) Inception module with dimension reductions

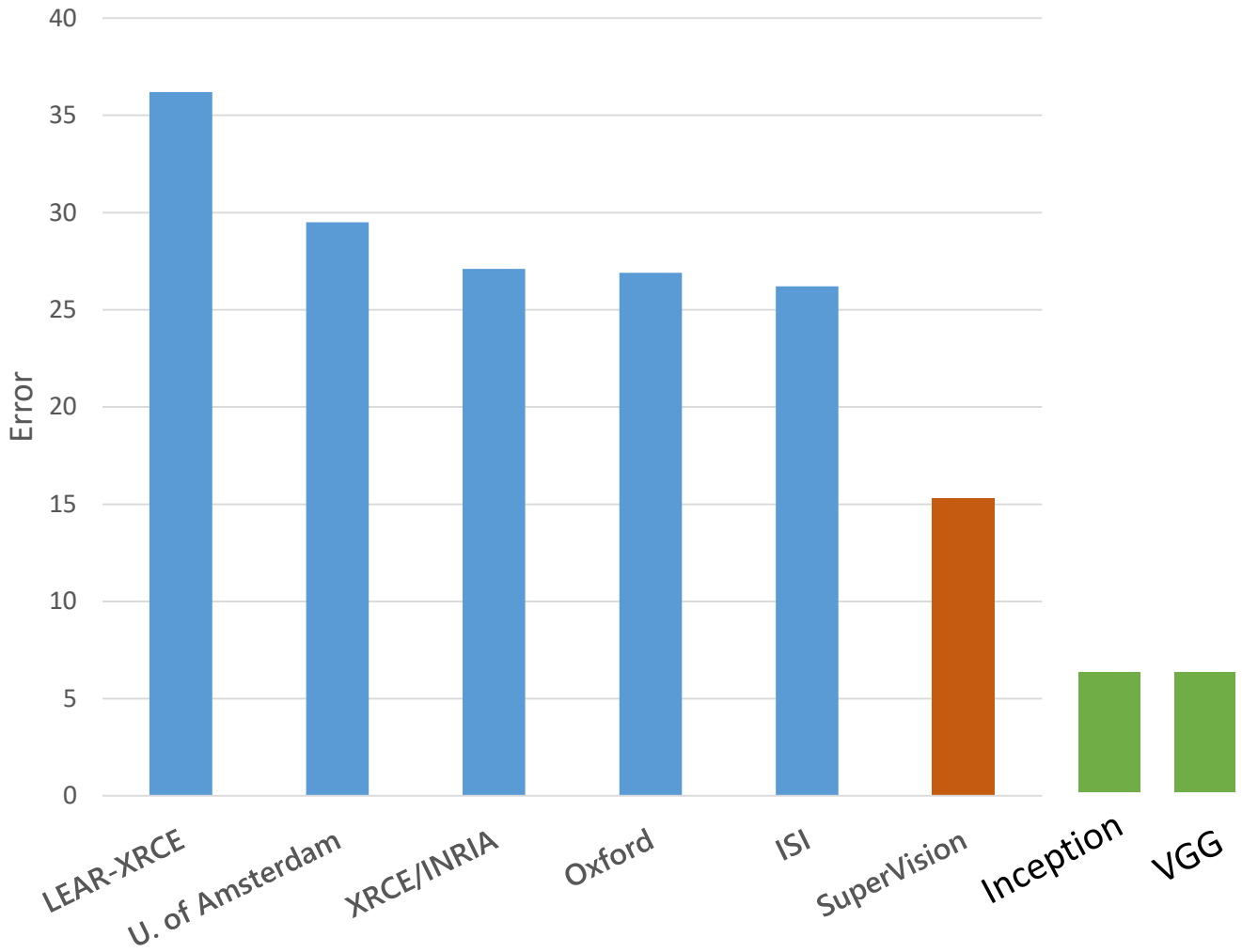
Going deeper with convolutions,
Szegedy, Liu, Jia, Sermanet, Reed, Anguelov, Erhan, Vanhoucke, Rabinovich, CVPR 2015

Going deeper…



2012 ImageNet 1K

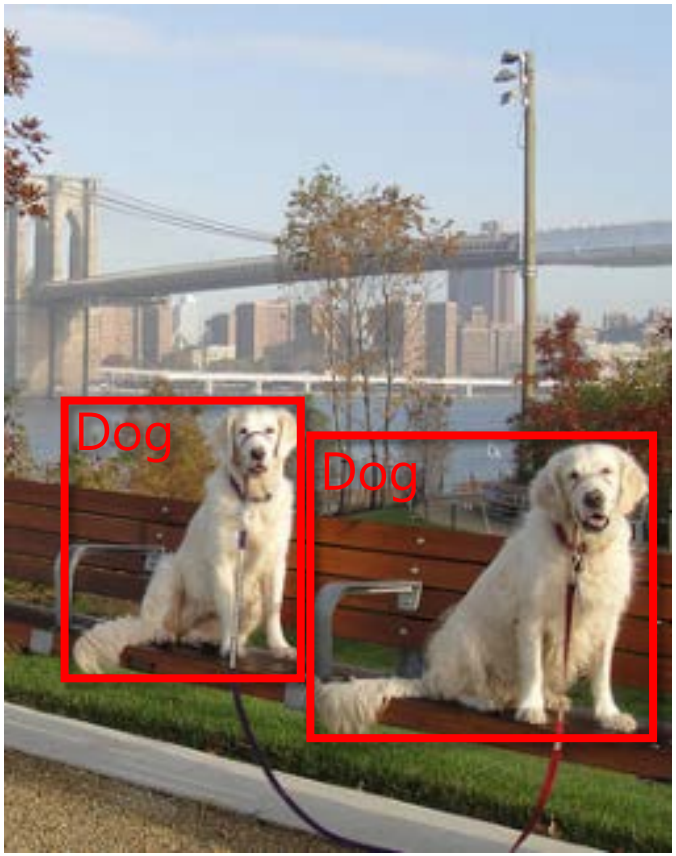
(Fall 2012)



2014 CNN vs. Detection

Classification vs. Detection

<

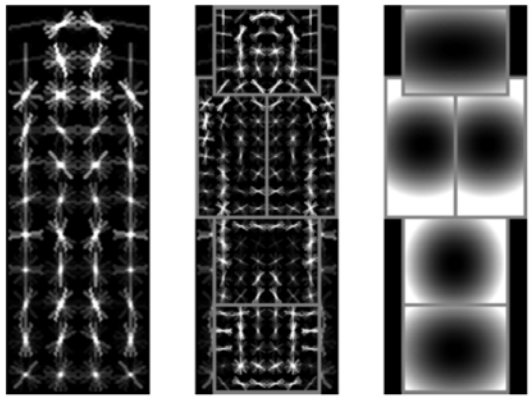


Let's try CNNs for detection, I think it might work.

Ross Girshick

Hmm, I doubt it.

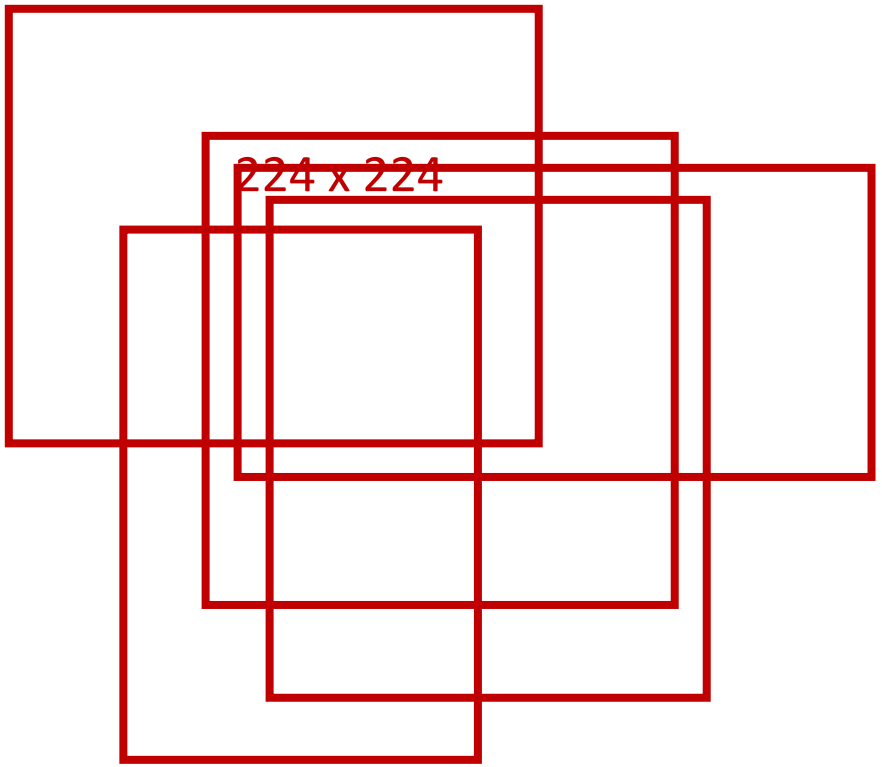
Jitendra Malik



DPM

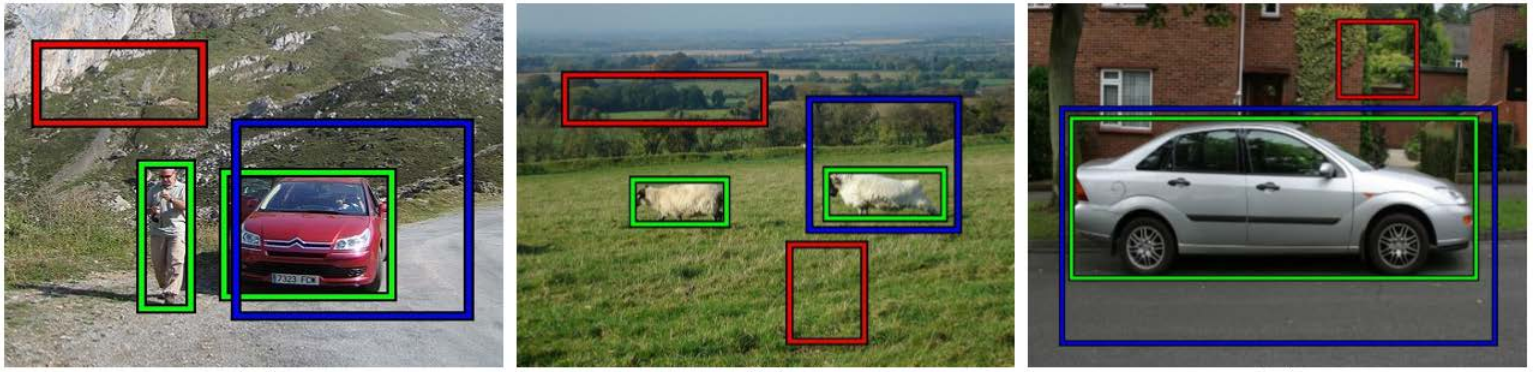
Aspect ratios?

DNNs are slow… (relatively)



Sliding window?

2010 Objectness

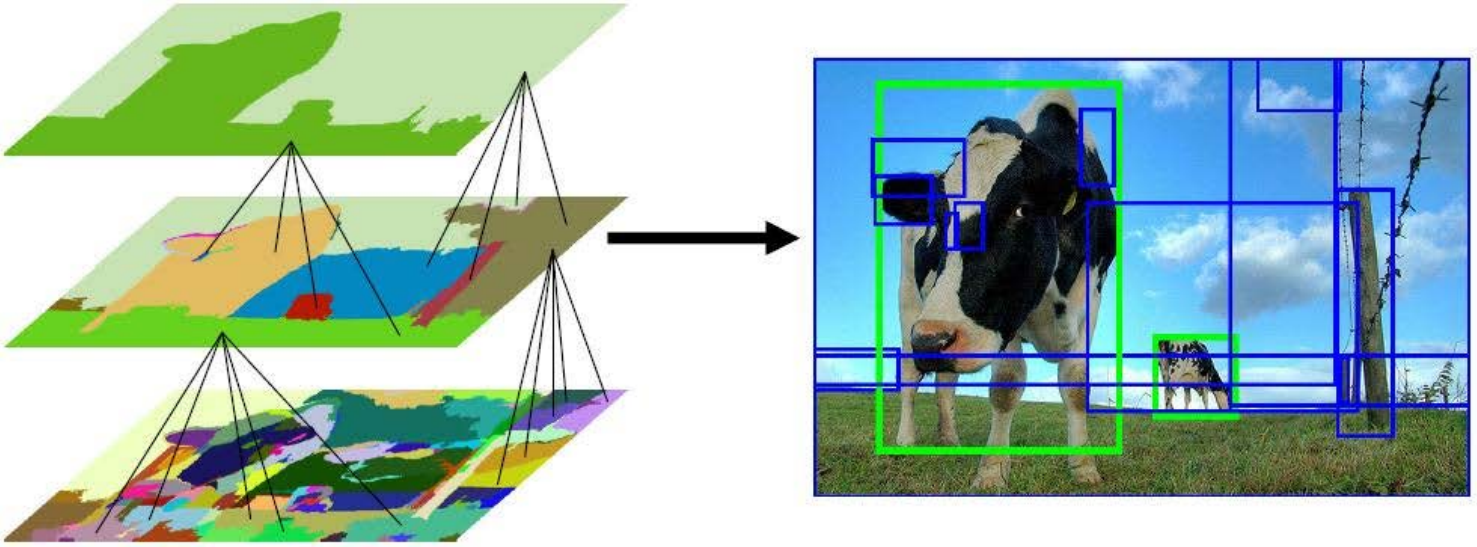


(a) (b) (c)

Fig. 1: **Desired behavior of an objectness measure.** *The desired objectness measure should score the blue windows, partially covering the objects, lower than the ground truth windows (green), and score even lower the red windows containing only stuff or small parts of objects.*

What is an object?, Bogdan Alexe, Thomas Deselaers, Vittorio Ferrari, CVPR 2010

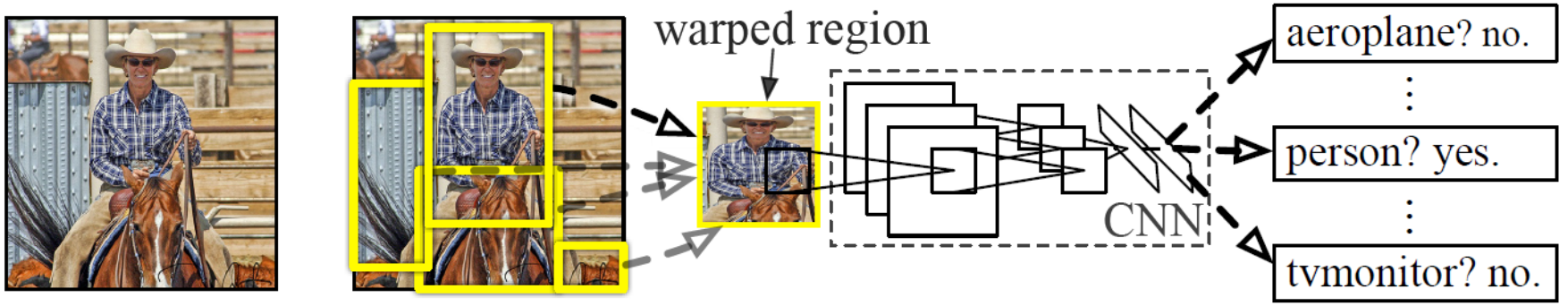
Object proposals



Selective Search for Object Recognition, Uijlings et al., 2013

Edge Boxes: Locating Object Proposals from Edges, Zitnick and Dollar, 2013

Object detection



1. Input image

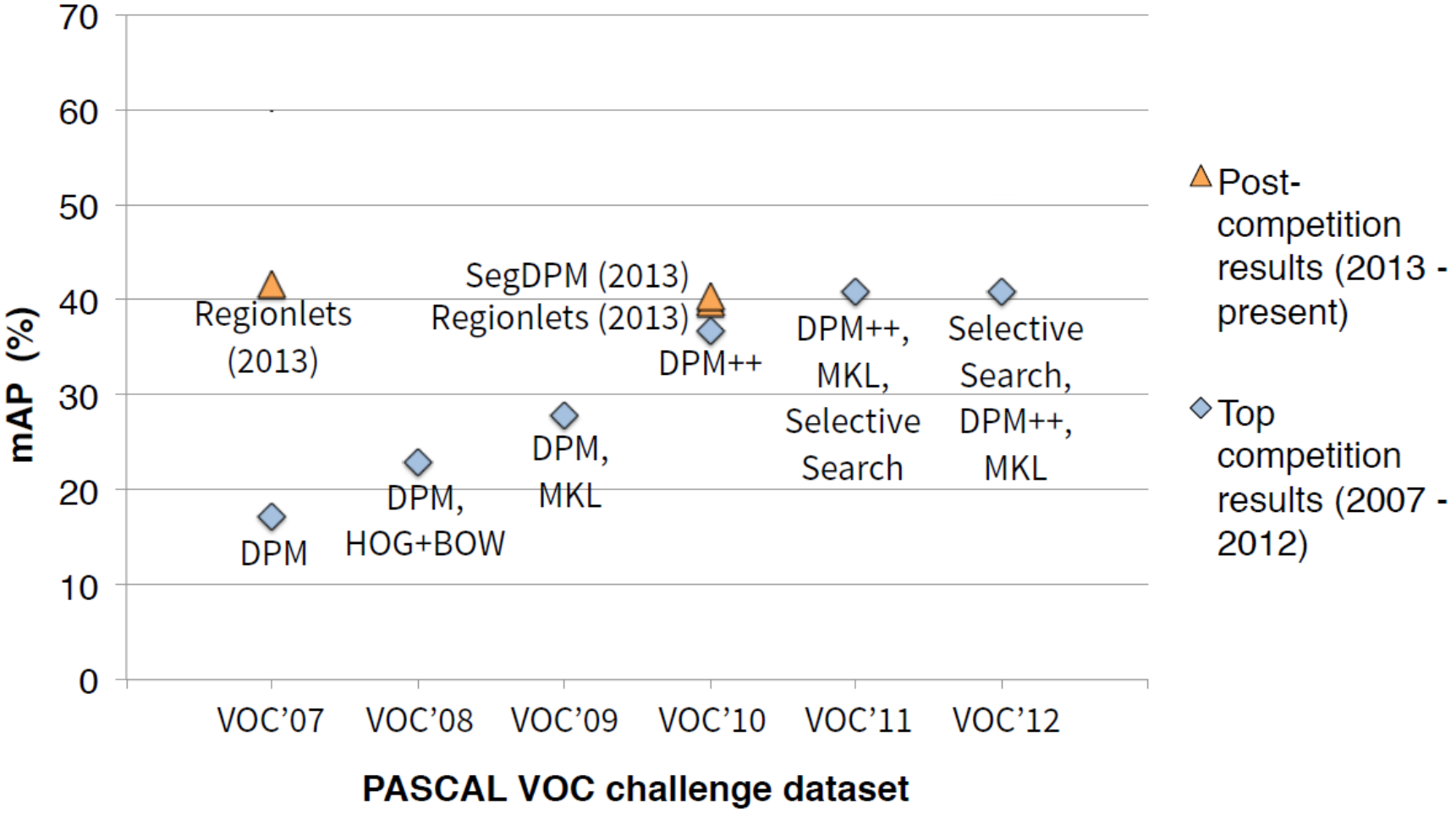
2. Extract region proposals (~2k)

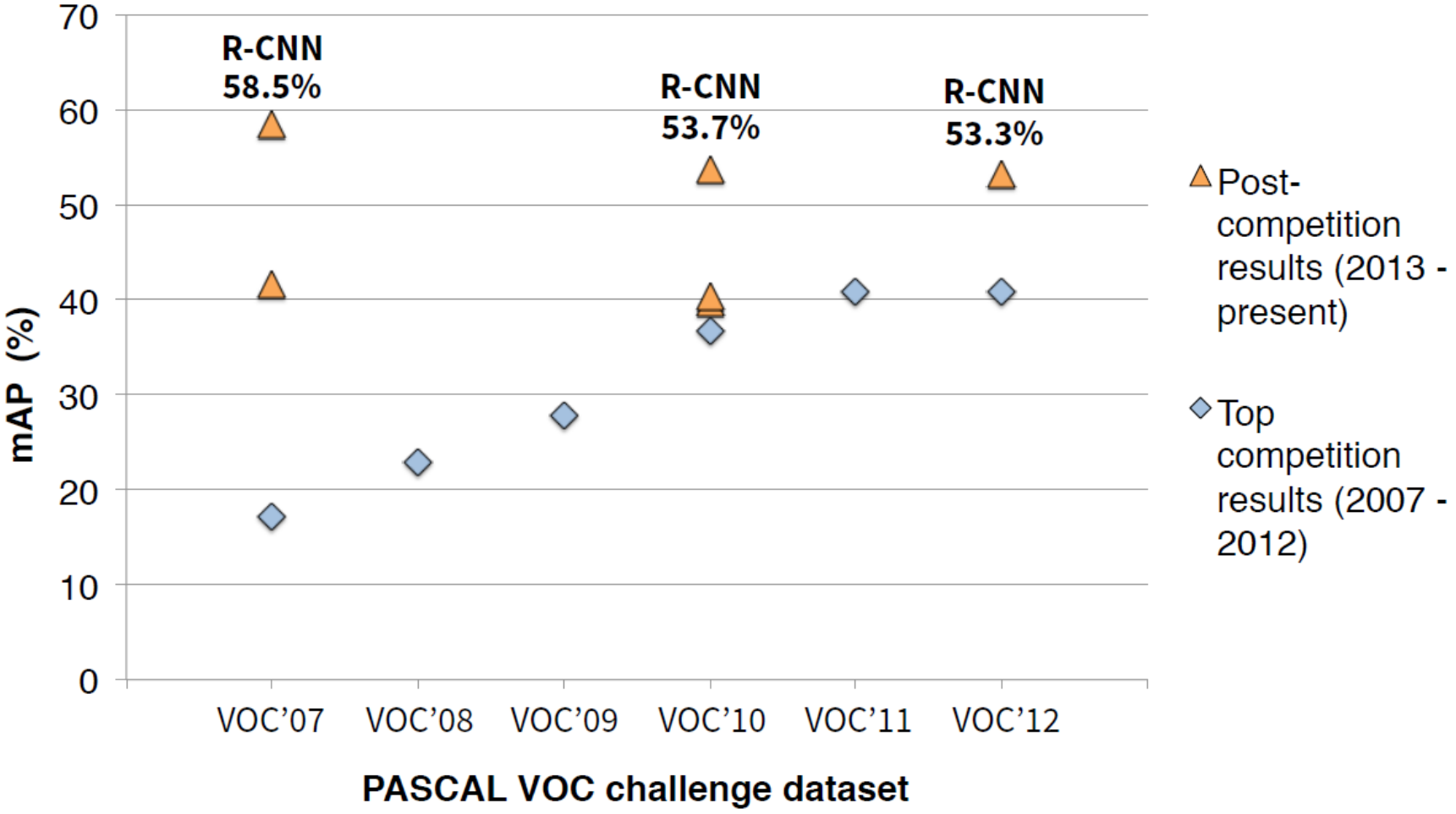
3. Compute CNN features

4. Classify regions

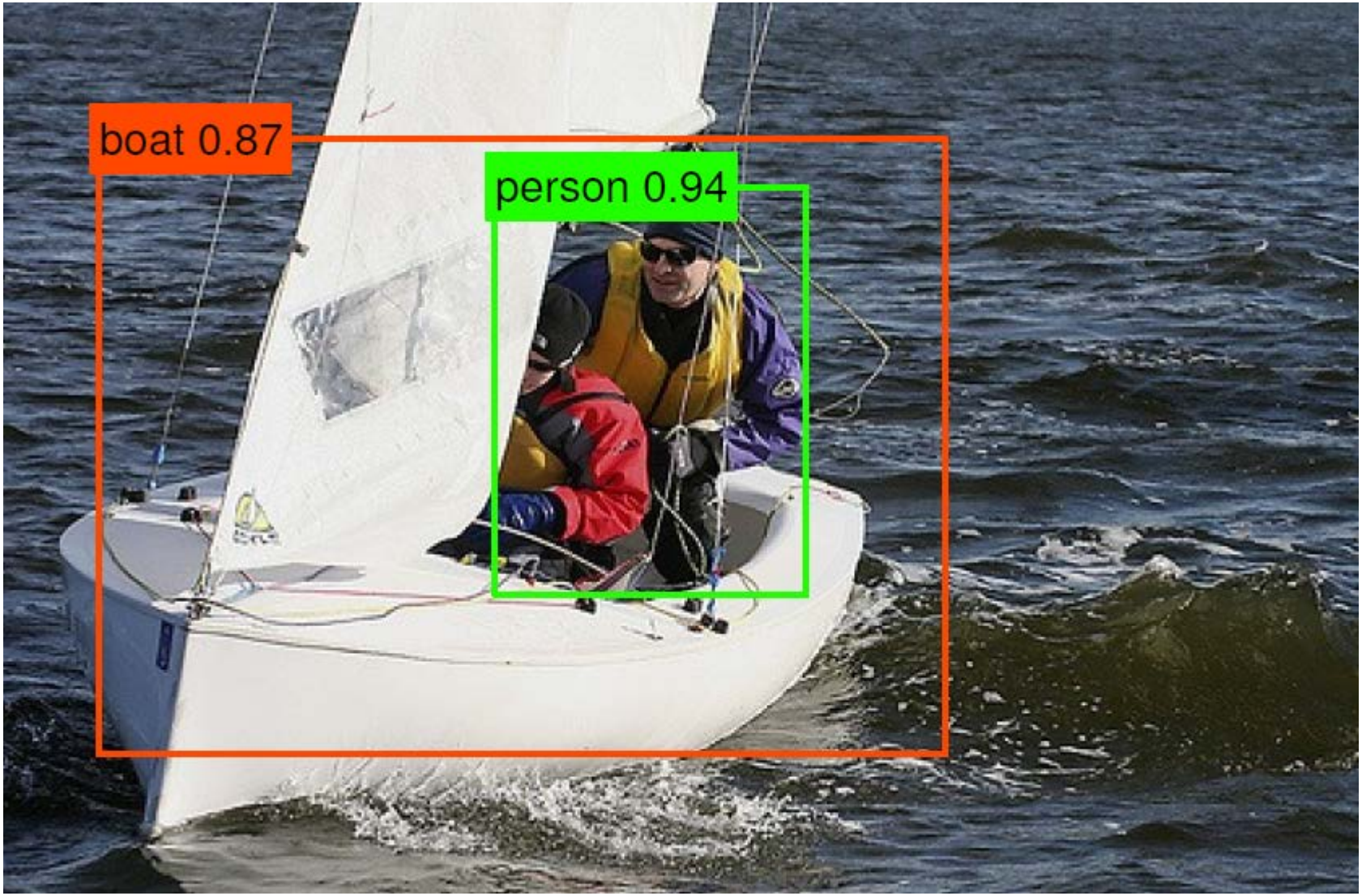
Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, Girshick, Donahue, Darrell, Malik, *CVPR* 2014.


TF07 WTS

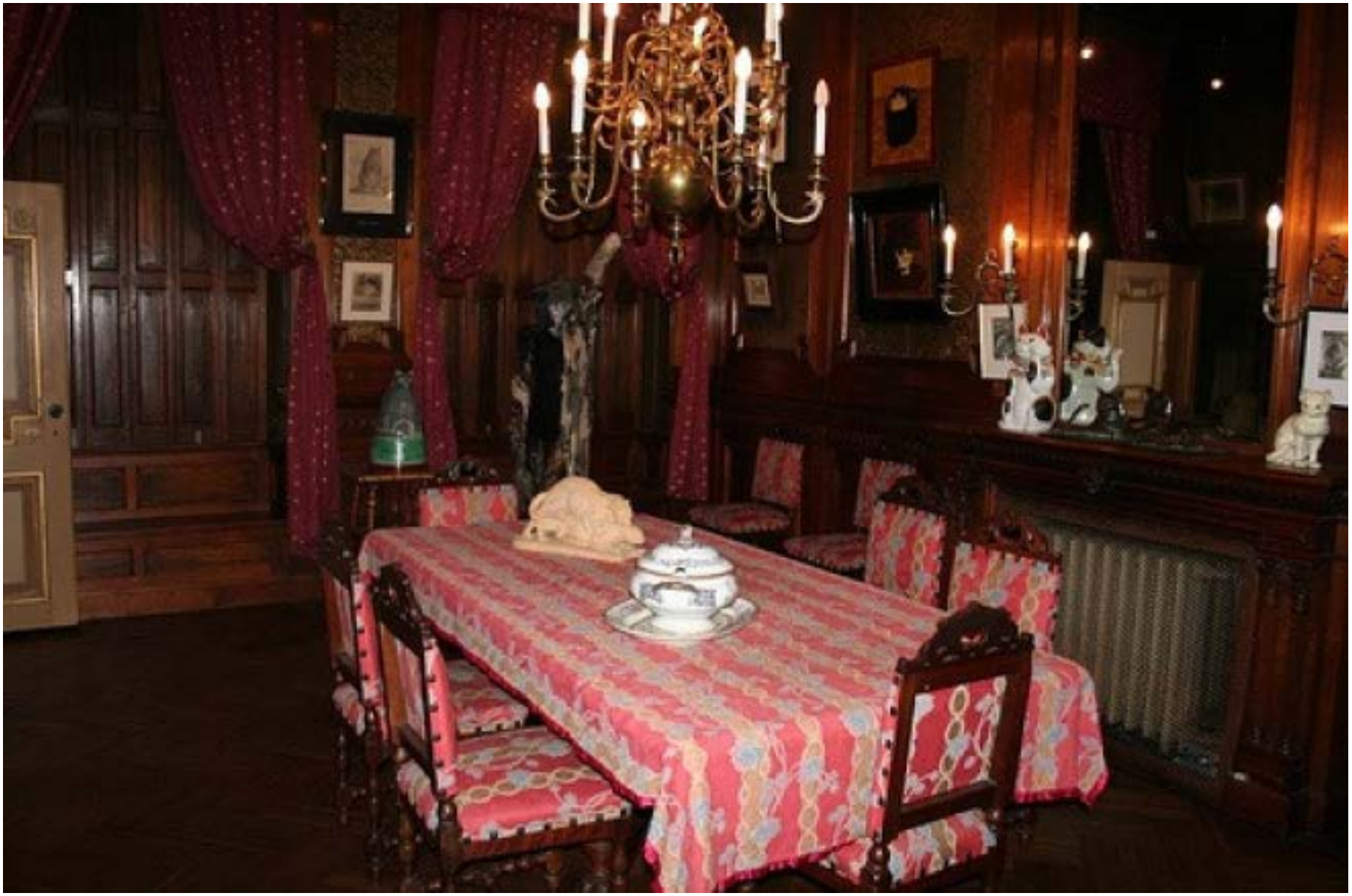

70
60
50
40
30
20
10
0
mAP (%)
Regionlets (2013)
DPM
DPM, HOG+BOW
DPM, MKL
SegDPM (2013)
Regionlets (2013)
DPM++
DPM++, MKL, Selective Search
Selective Search, DPM++, MKL
VOC'07
VOC'08
VOC'09
VOC'10
VOC'11
VOC'12
PASCAL VOC challenge dataset
Post-competition results (2013 - present)
Top competition results (2007 - 2012)

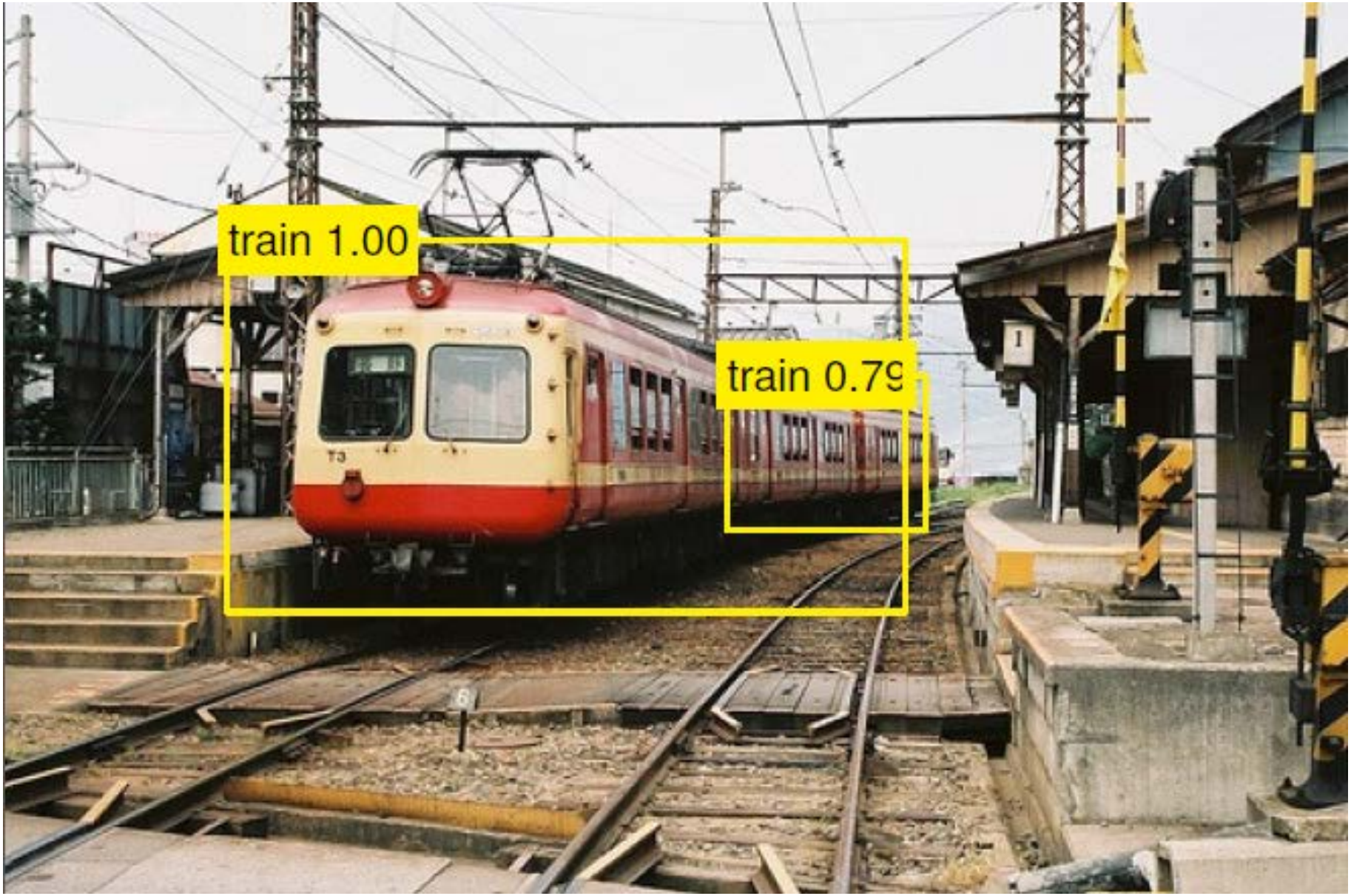

70
60
50
40
30
20
10
0
mAP (%)
R-CNN
58.5%
R-CNN
53.7%
R-CNN
53.3%
VOC'07
VOC'08
VOC'09
VOC'10
VOC'11
VOC'12
PASCAL VOC challenge dataset
Post-competition results (2013 - present)
Top competition results (2007 - 2012)

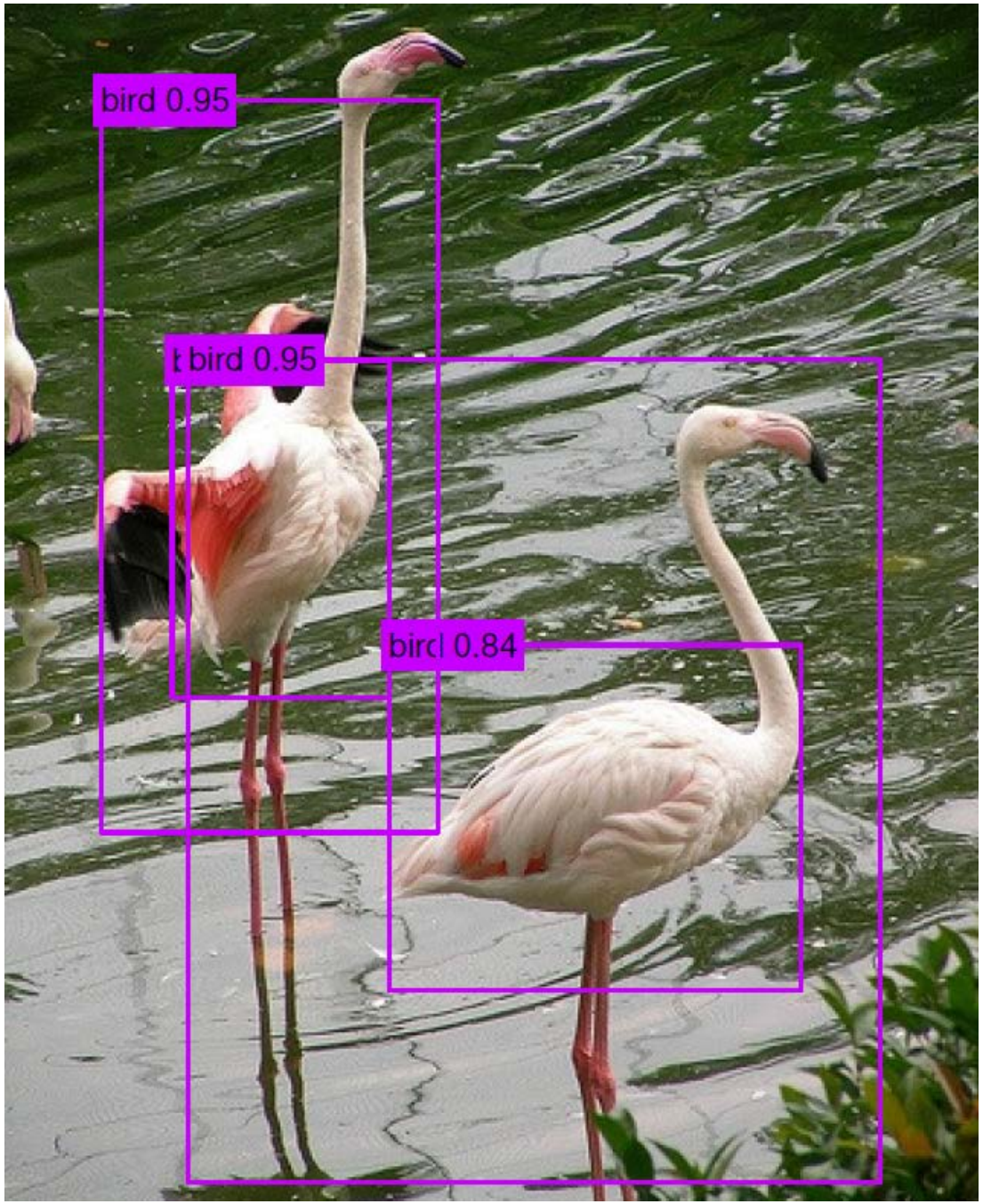

train 1.00

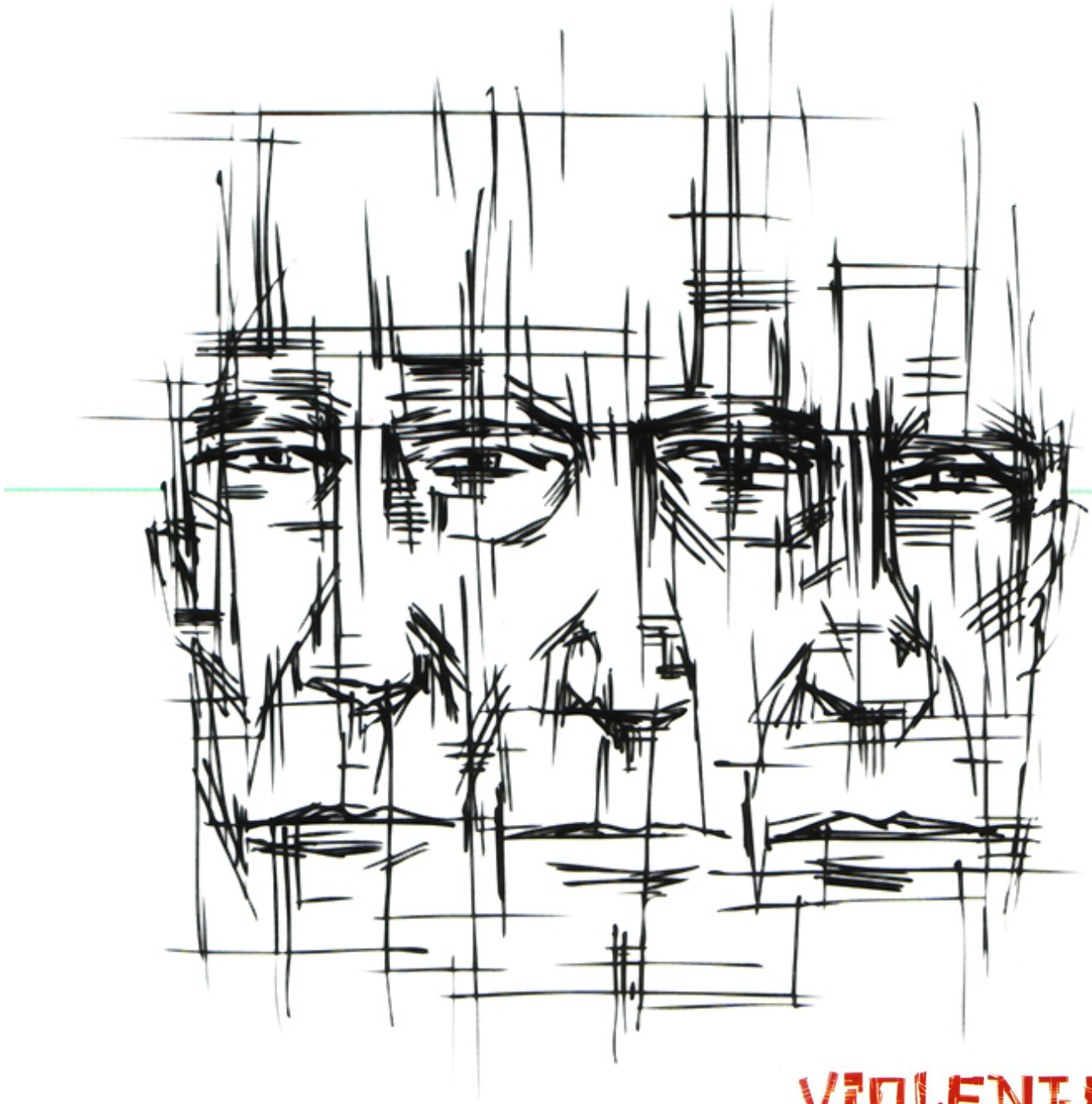

boat 0.87
person 0.94


bottle 0.98
bottle 0.86
bottle 0.82




train 1.00
train 0.79


bird 0.95
t bird 0.95
bird 0.84


VIOLENT FEMMES


cat 0.95


cat 0.82
http://phyllischan.blogspot.com

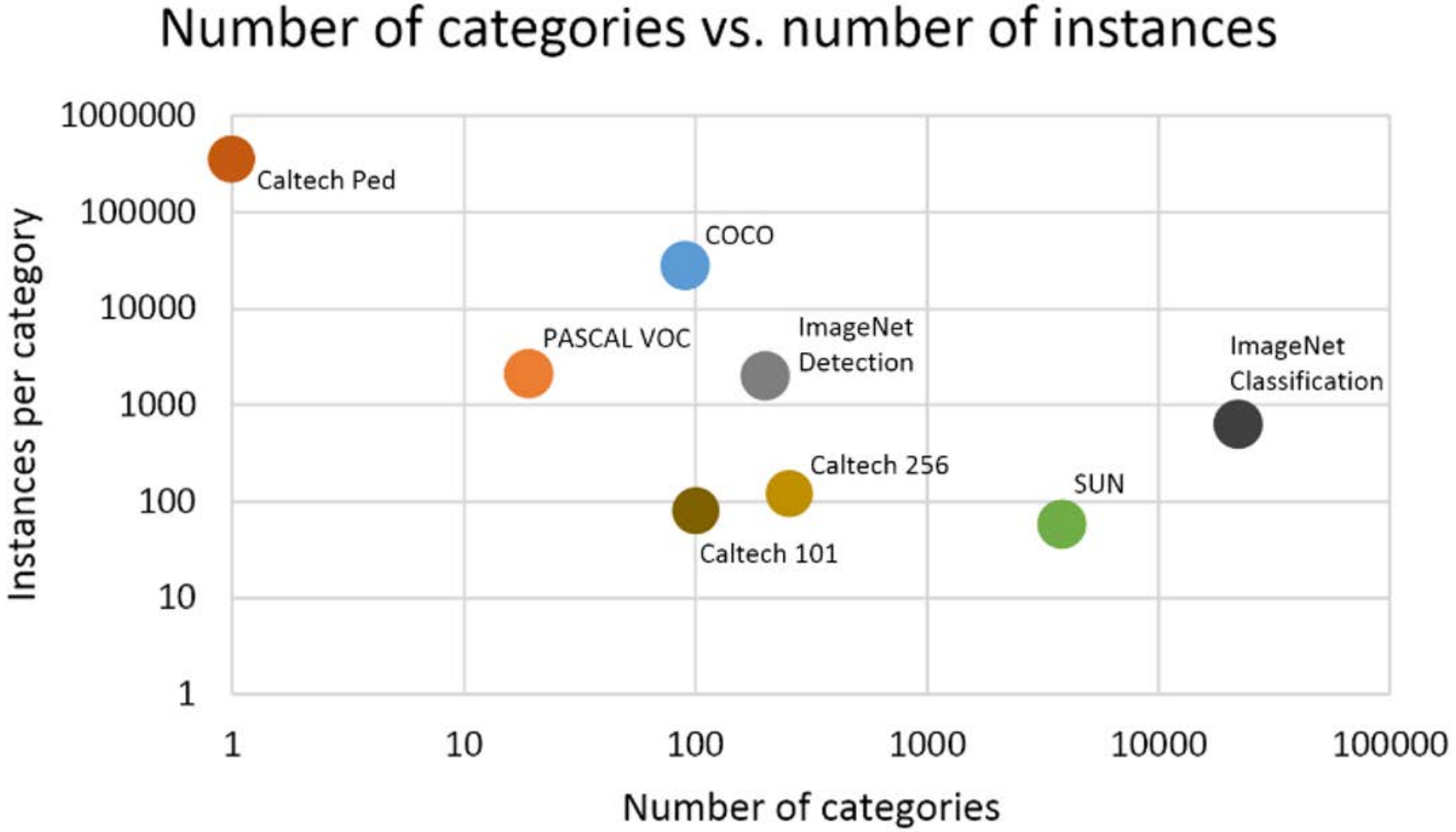

person 0.52
dumbbell 1.00

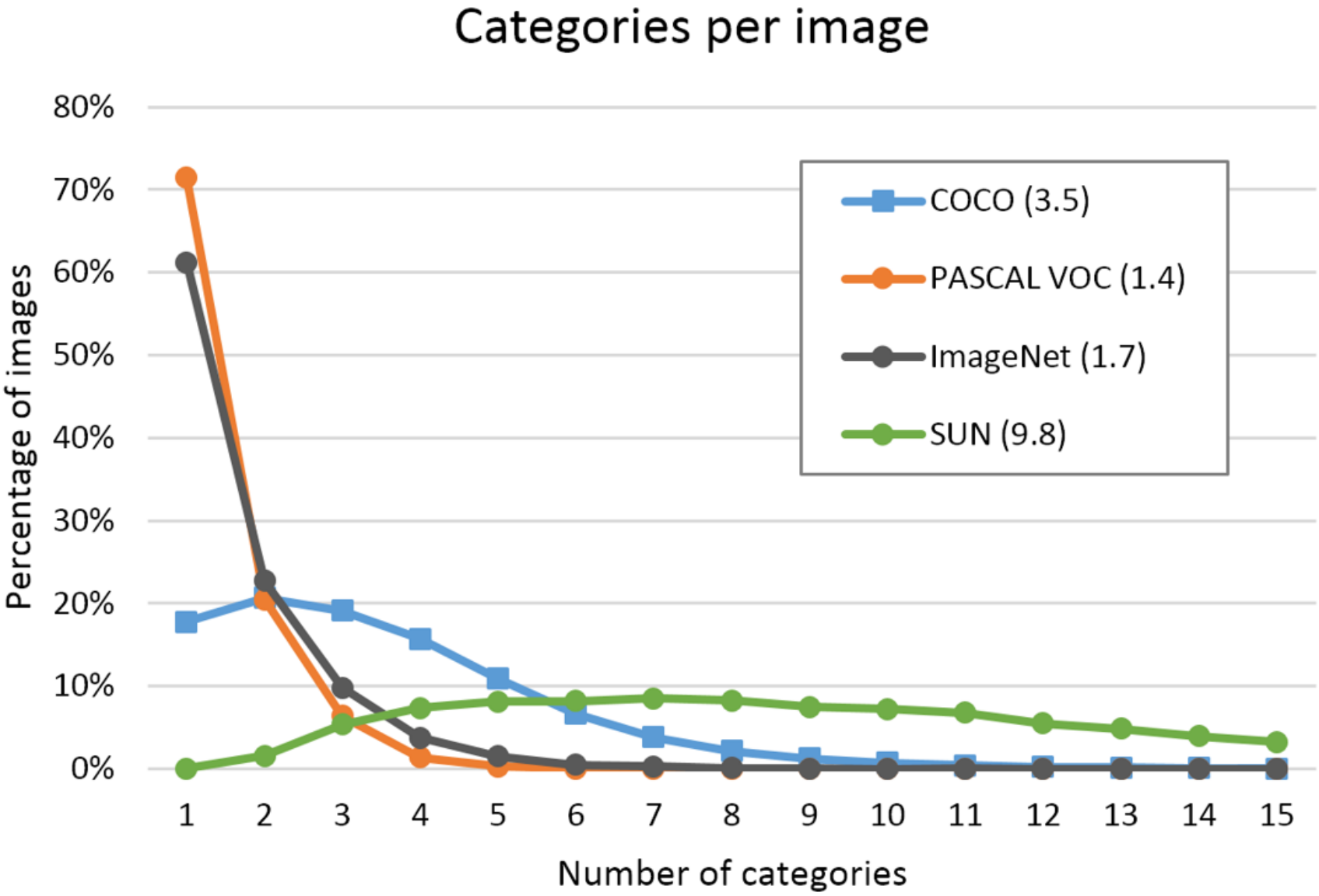
http://cocodataset.org

Microsoft COCO: Common Objects in Context,
Lin, Maire, Belongie, Hays, Perona, Ramanan,
Dollár, Zitnick, *ECCV* 2014.

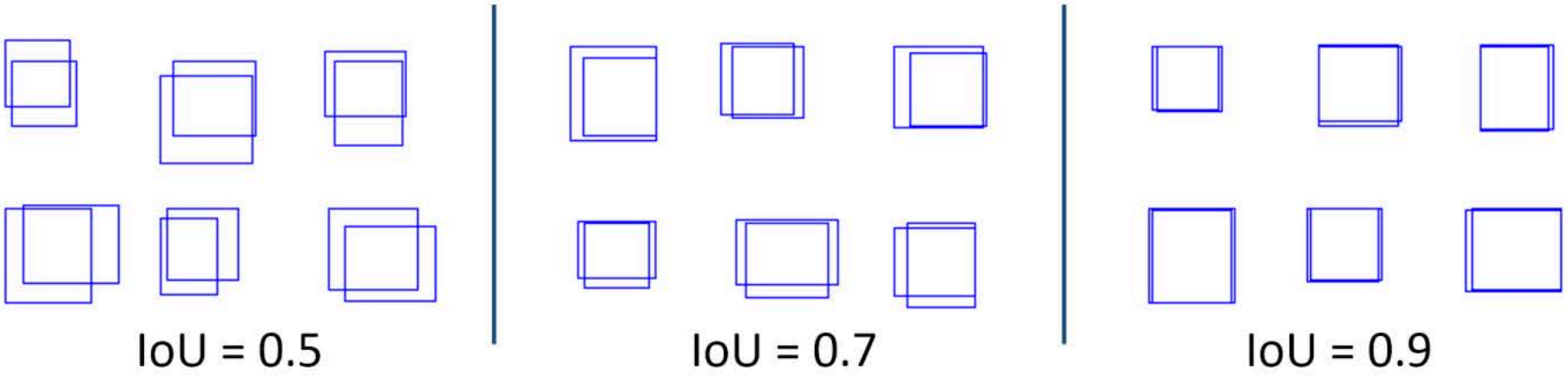
Iconic object images

Non-iconic images


Number of categories vs. number of instances
Instances per category
1000000
100000
10000
1000
100
10
1
Caltech Ped
COCO
PASCAL VOC
ImageNet Detection
ImageNet Classification
Caltech 256
SUN
Caltech 101
1
10
100
1000
10000
100000
Number of categories


Categories per image
80%
70%
60%
50%
40%
30%
20%
10%
0%
Percentage of images
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Number of categories
COCO (3.5)
PASCAL VOC (1.4)
ImageNet (1.7)
SUN (9.8)

Intersection over Union (IoU)

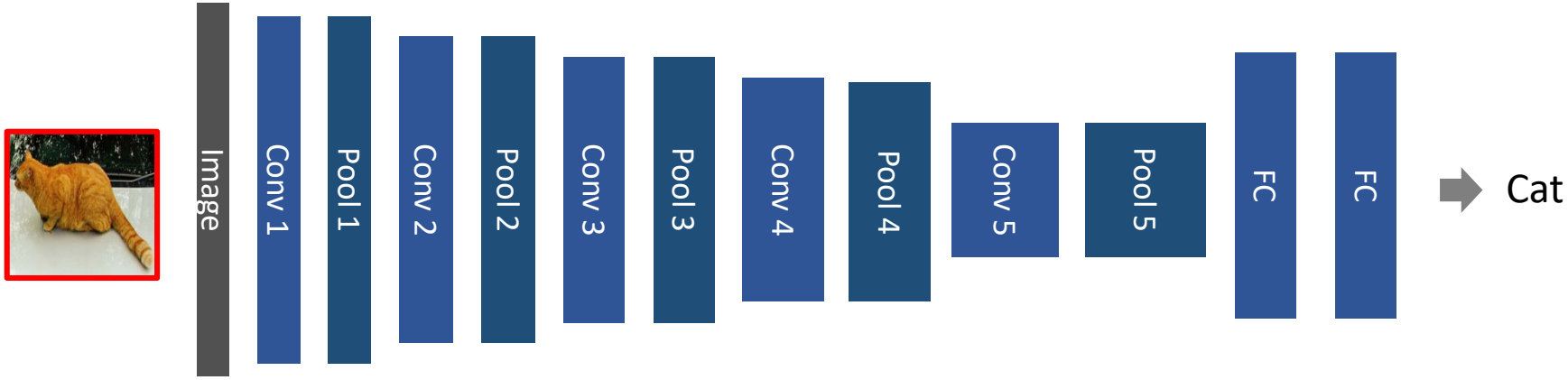


COCO

Over 55,000 worker hours (6+ years)

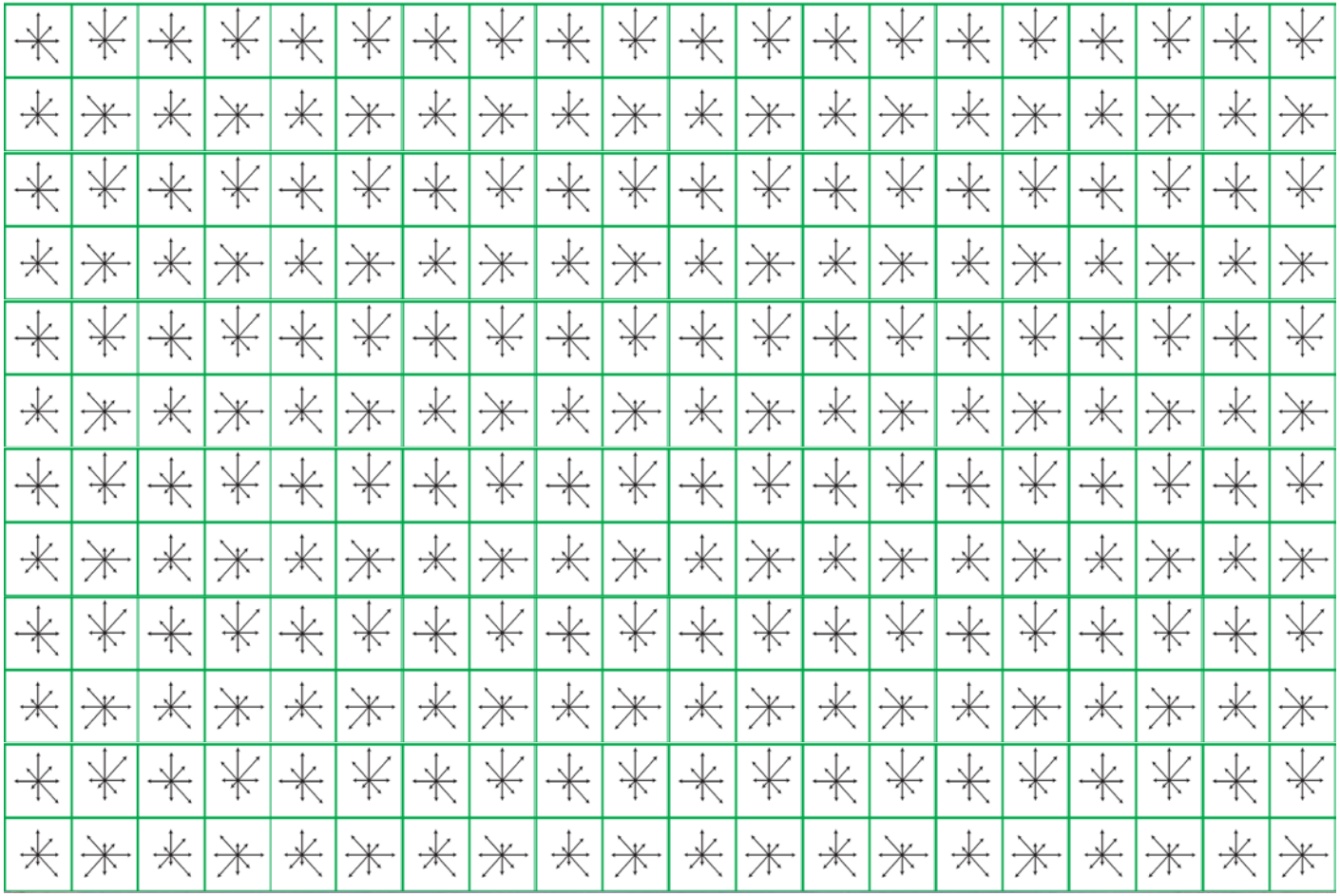
- 70-100 object categories (things not stuff)
- 330,000 images (~150k first release)
- 2 million instances (400k people)
- Every instance is segmented
- 7.7 instances per image (3.5 categories)
- Key points
- 5 sentences per image

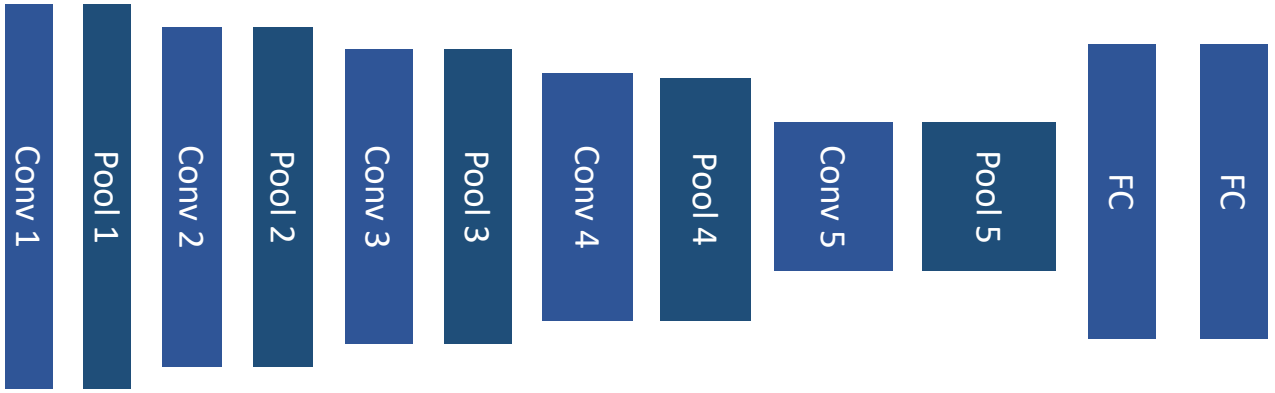
RCNN is slow…



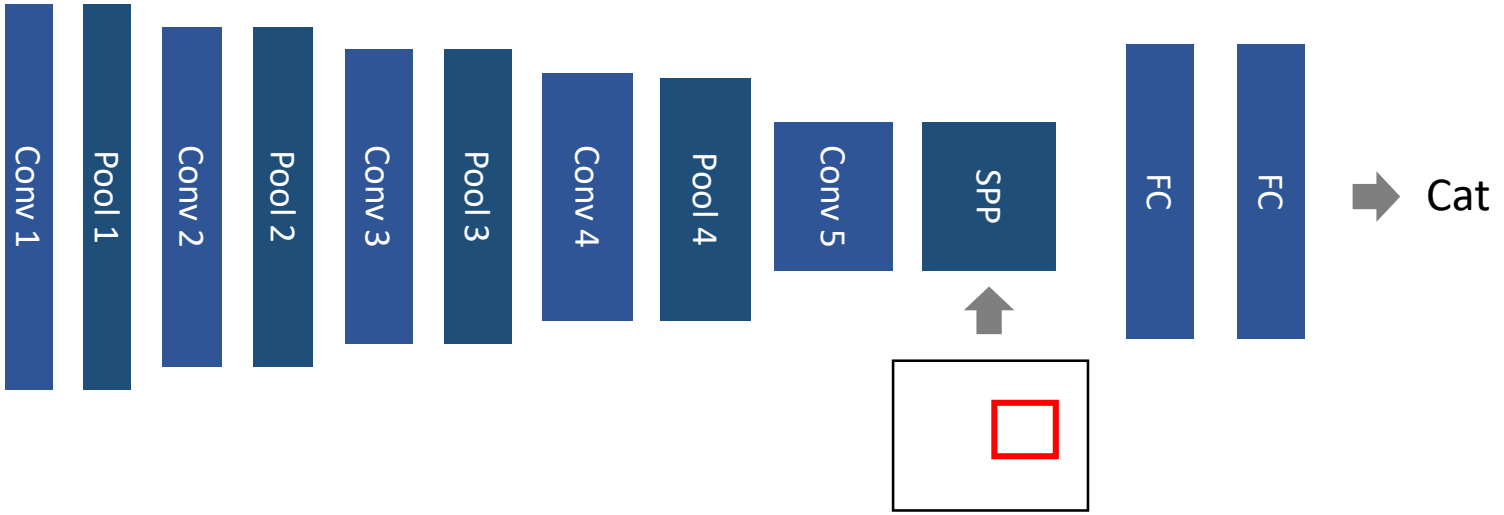
Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition
Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, ECCV 2014

2005 HoG (Dalal and Triggs)





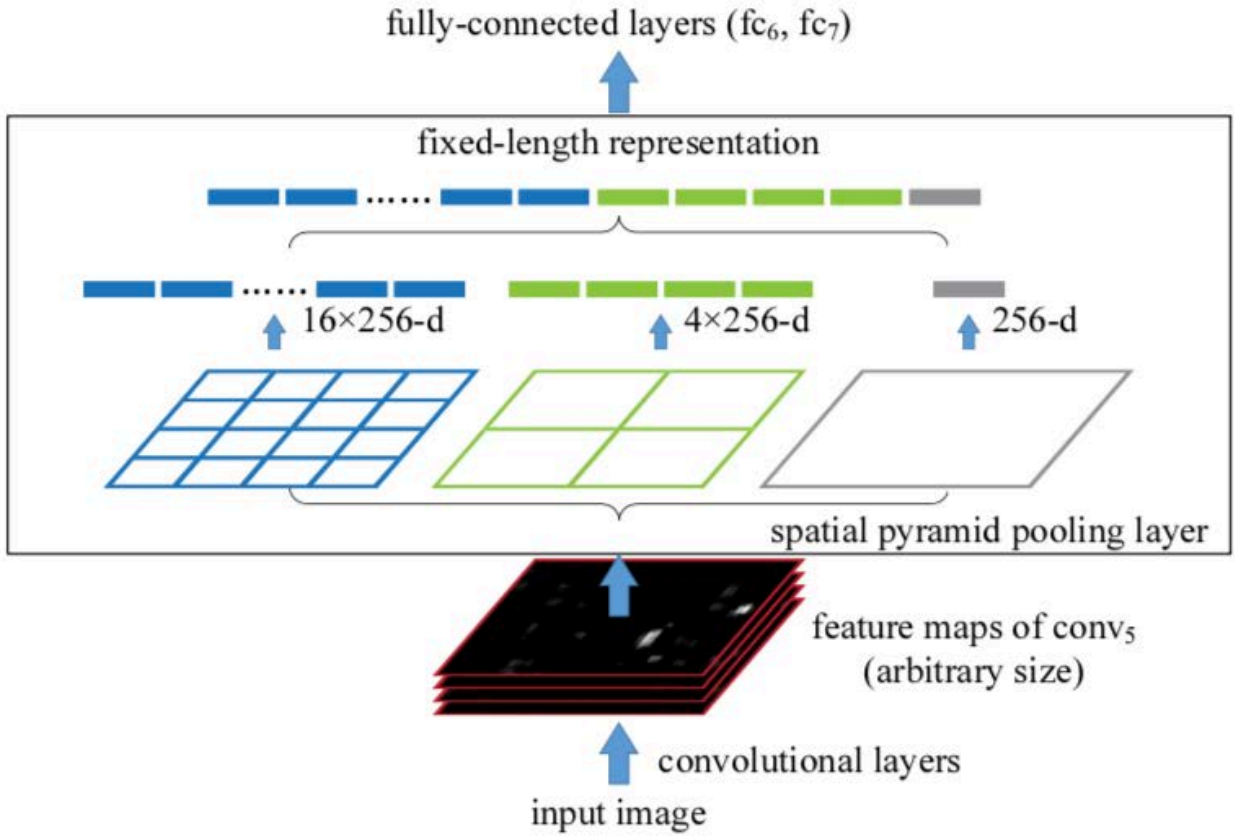
VS.



Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, ECCV 2014

Spatial pyramid pooling layer



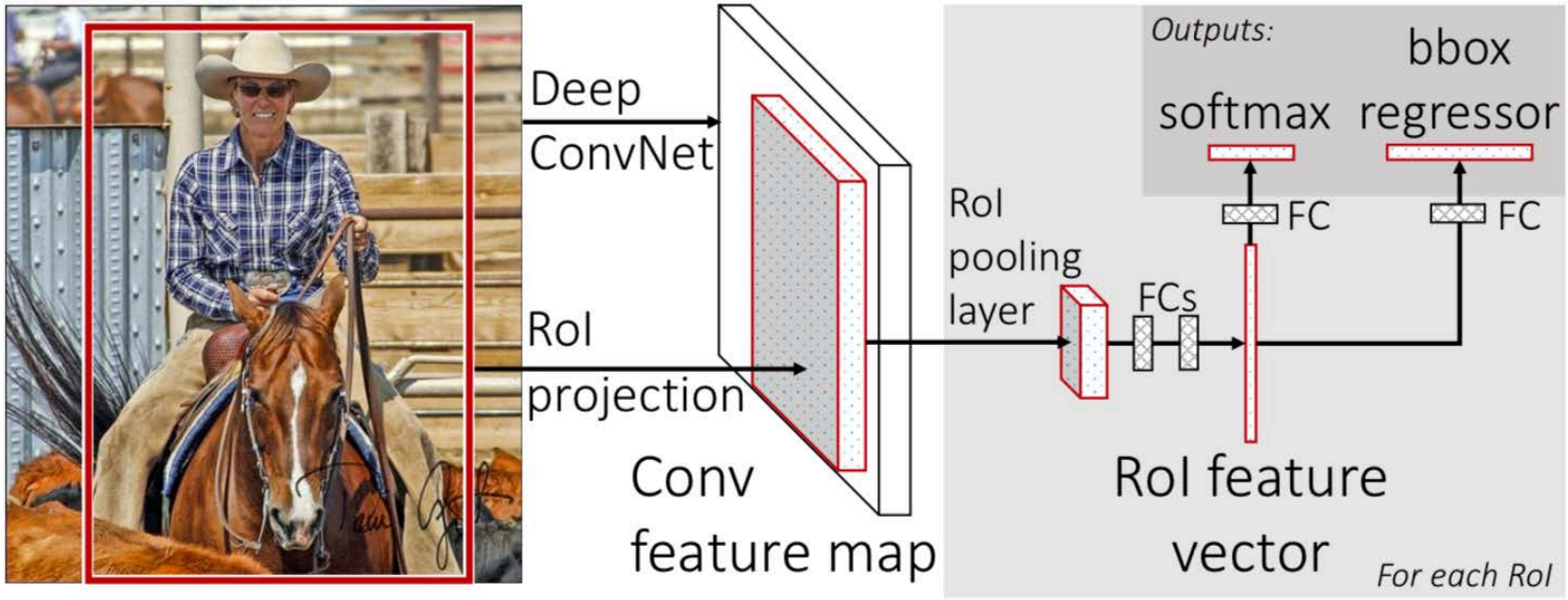
Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition
Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, ECCV 2014

Simplifying training…

R-CNN training pipeline:

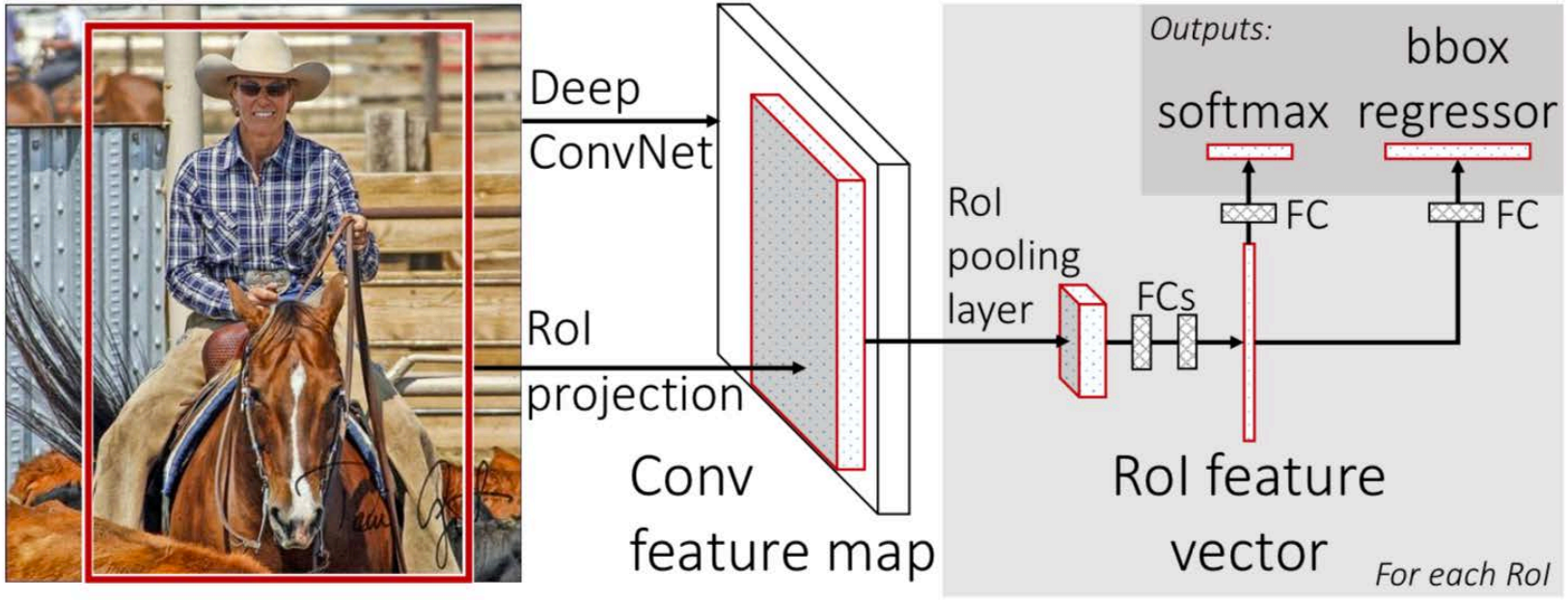
1) Fine tune a ConvNet on object proposals using log loss
2) Fit SVMs to ConvNet features
3) Learn bounding-box regressors

Simplifying training



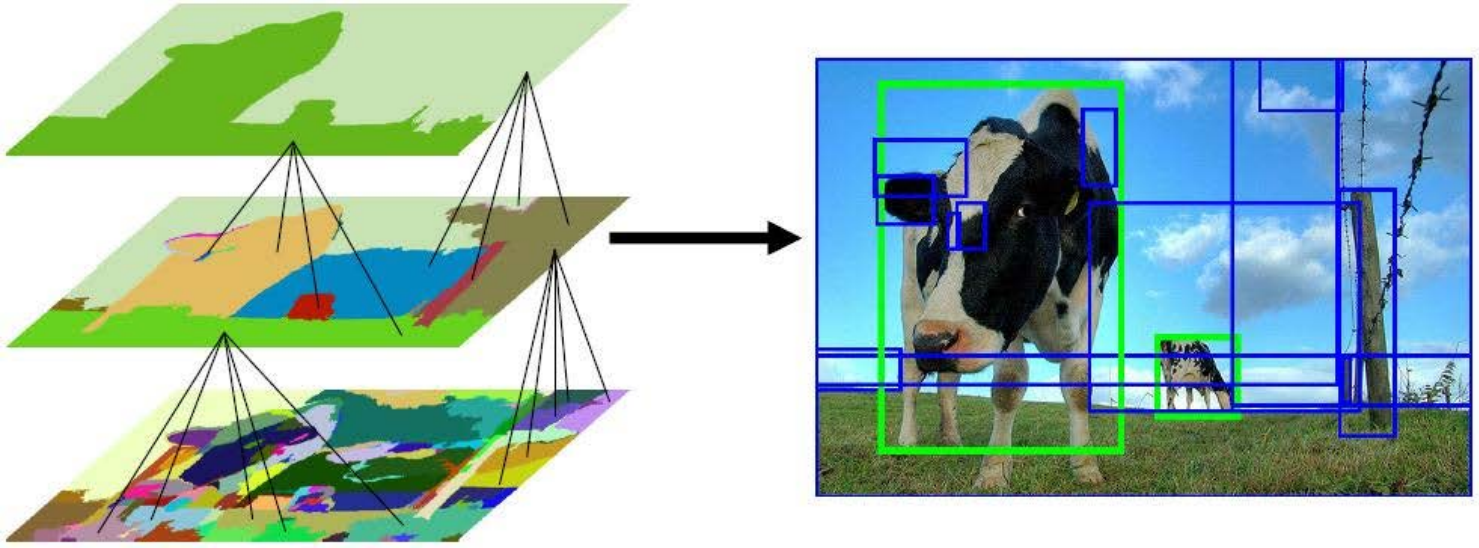
Fast R-CNN
Ross Girshick, ICCV 2015

What is the computational bottleneck?



Getting rid of all “handcrafting”

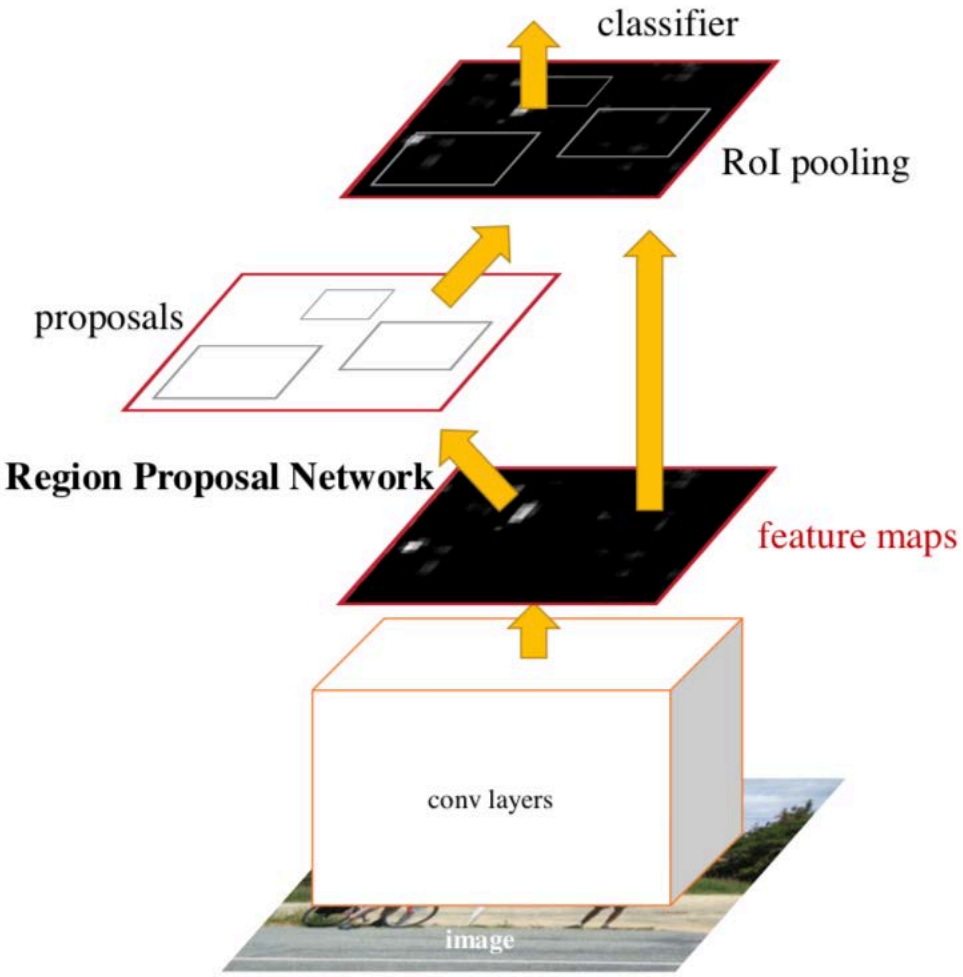
2 seconds per image in a CPU implementation



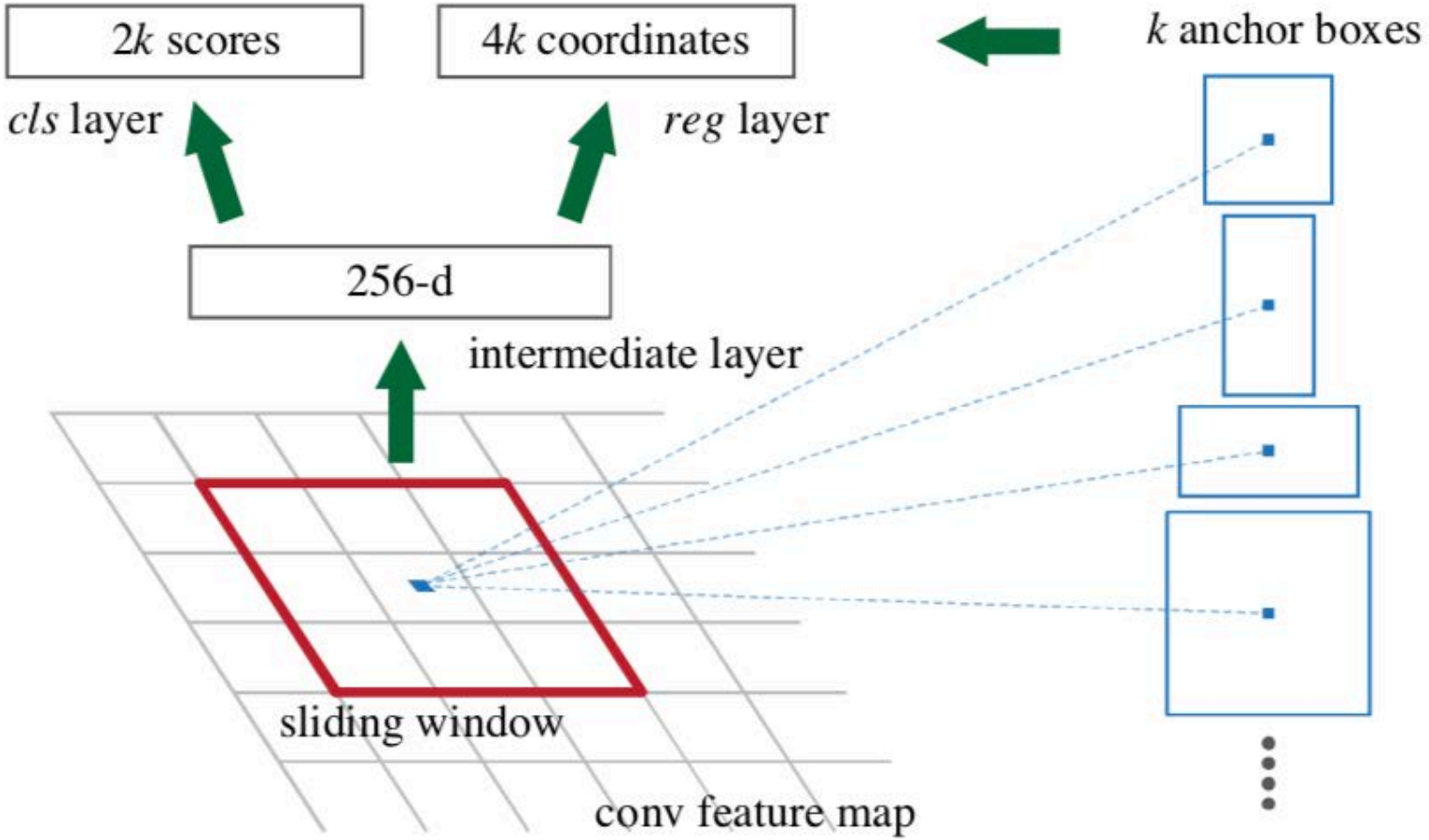
Selective Search for Object Recognition, Uijlings et al., 2013

Edge Boxes: Locating Object Proposals from Edges, Zitnick and Dollar, 2013

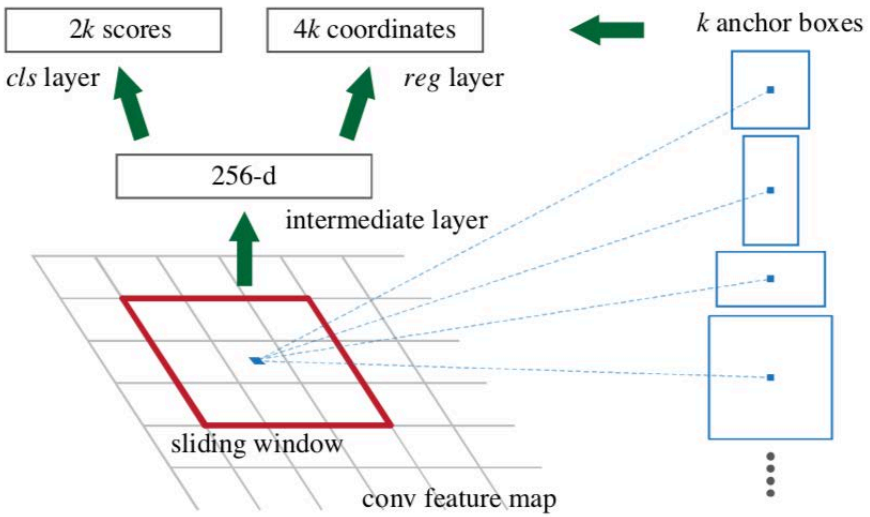
Learning to Refine Object Segments, Pinheiro, Lin, Collobert, Dollár, 2016



Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks
S Ren, K He, R Girshick, J Sun



Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks
S Ren, K He, R Girshick, J Sun



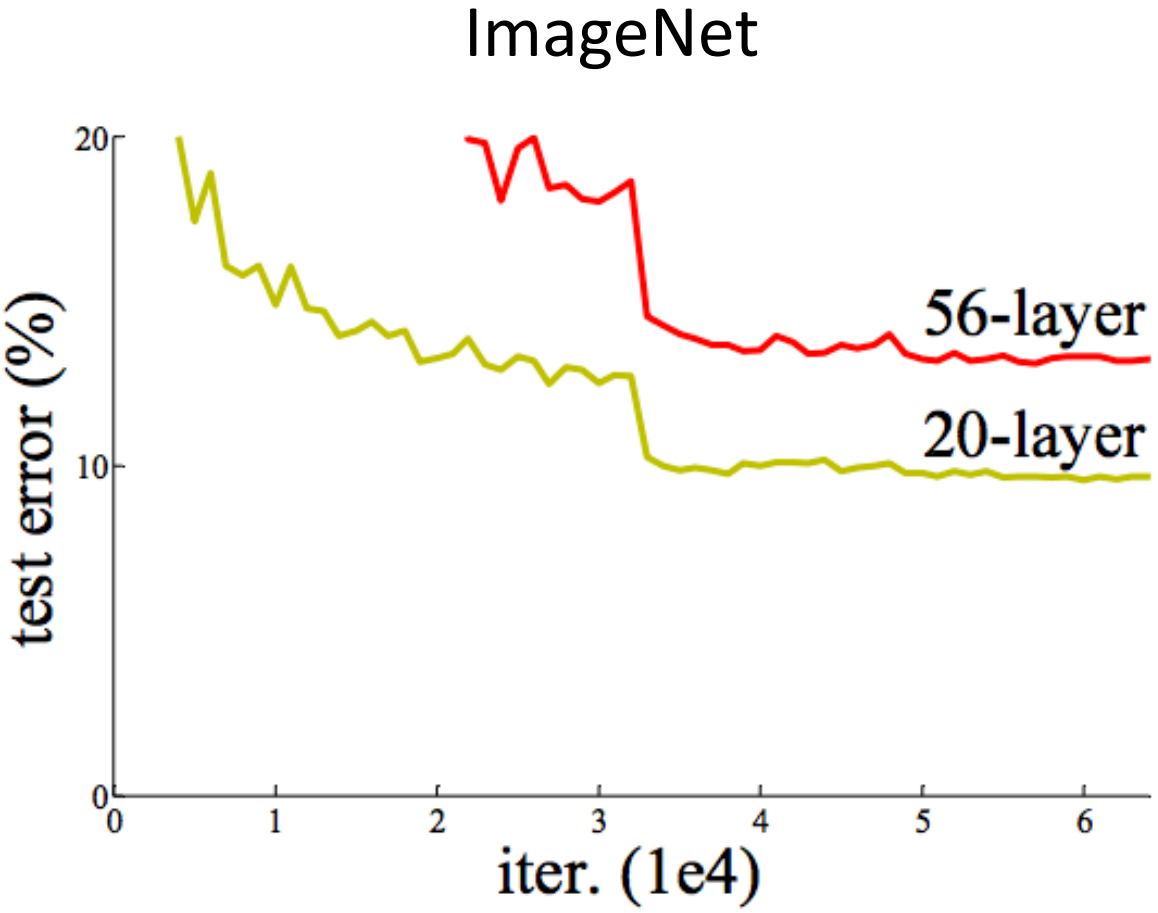
Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks
S Ren, K He, R Girshick, J Sun

COCO Object Detection Average Precision (%)

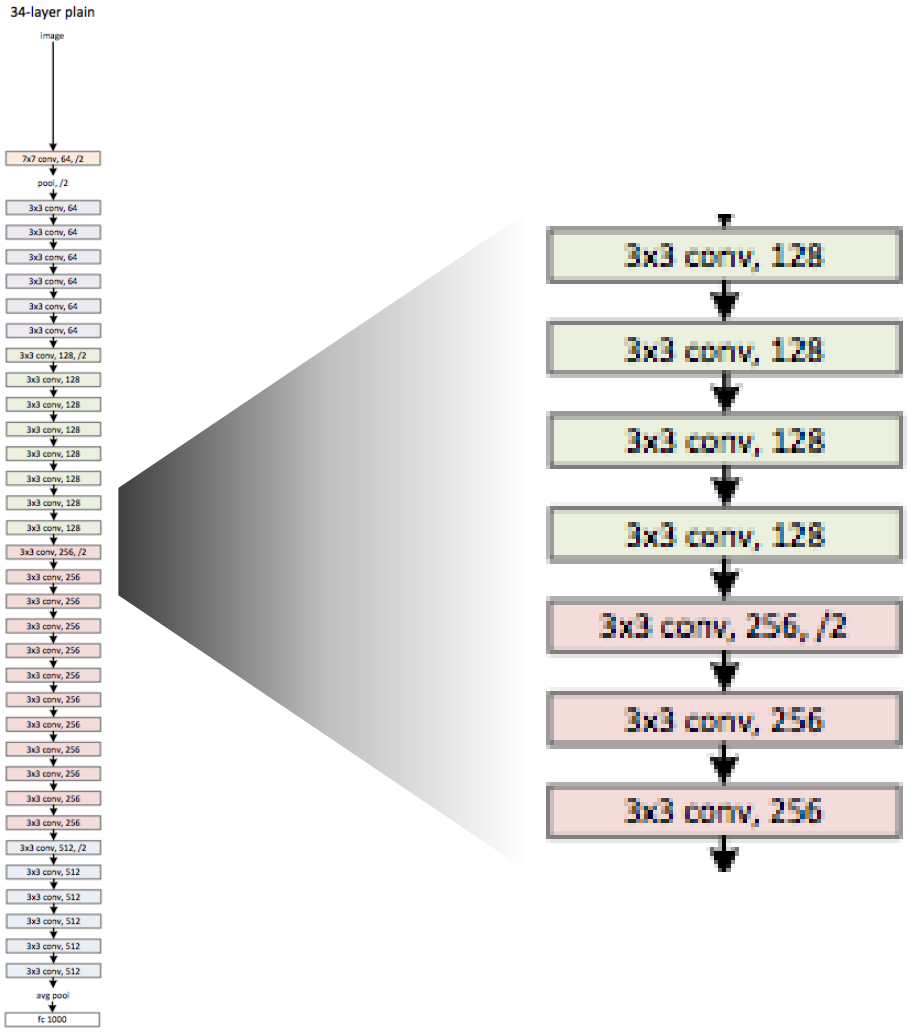
2006

How can we train deeper networks?

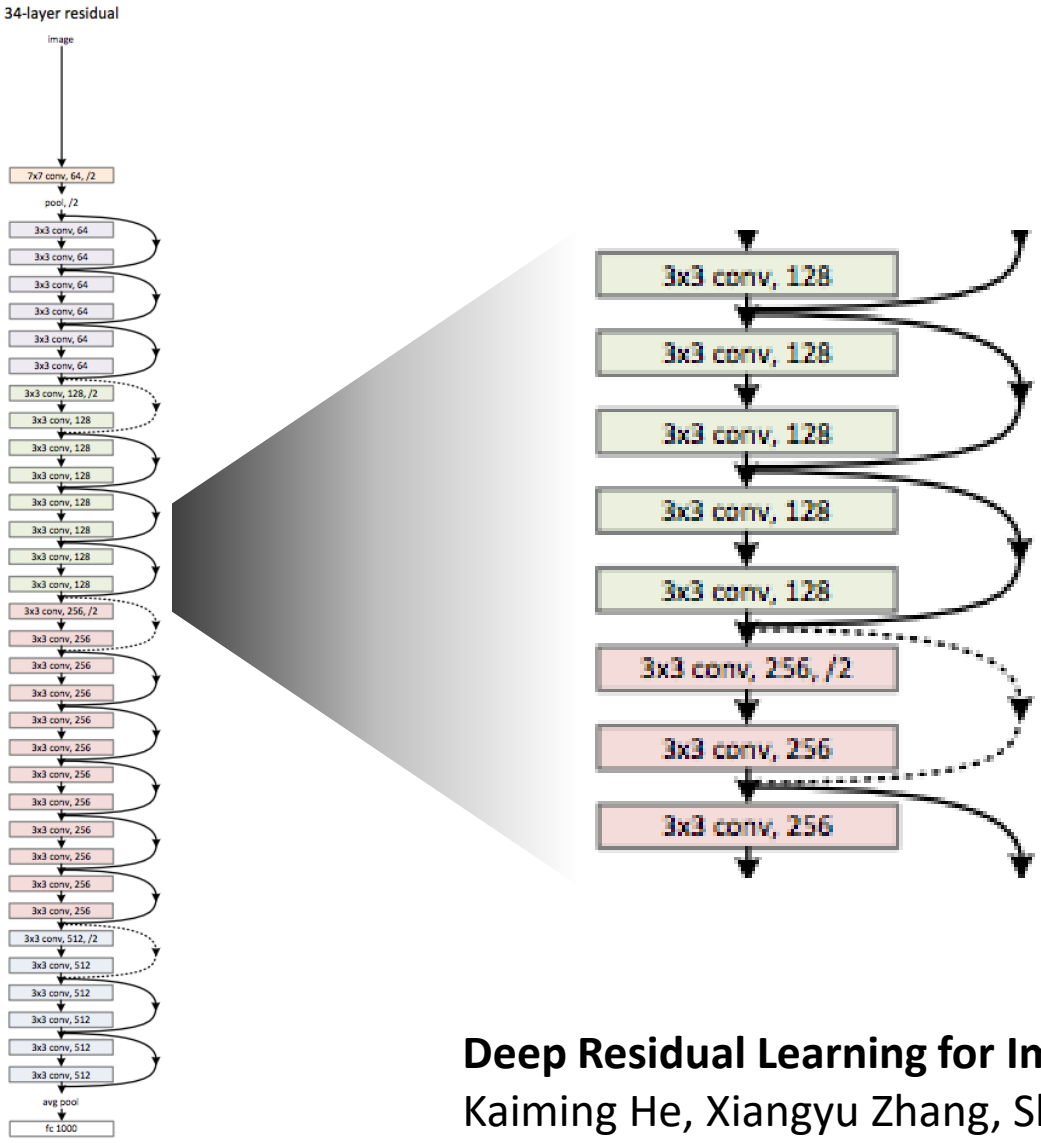
How do we even go deeper?



Plain deep networks



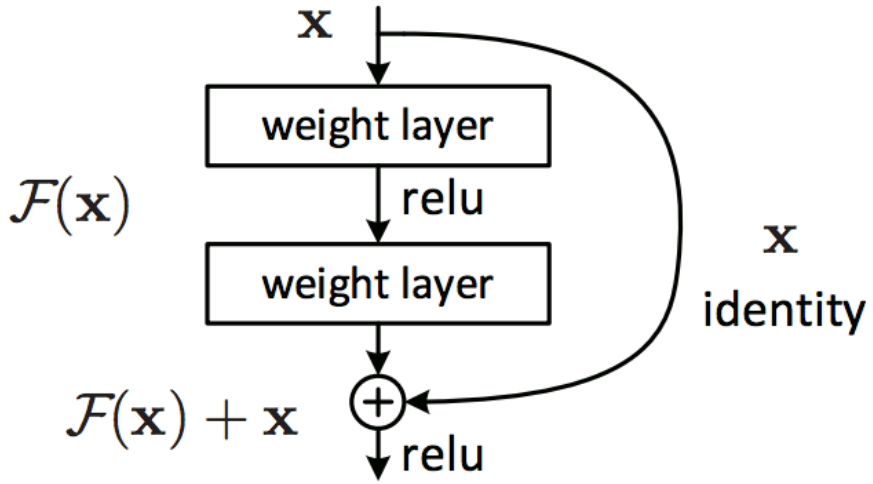
ResNets (residual networks)



Deep Residual Learning for Image Recognition

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, 2015

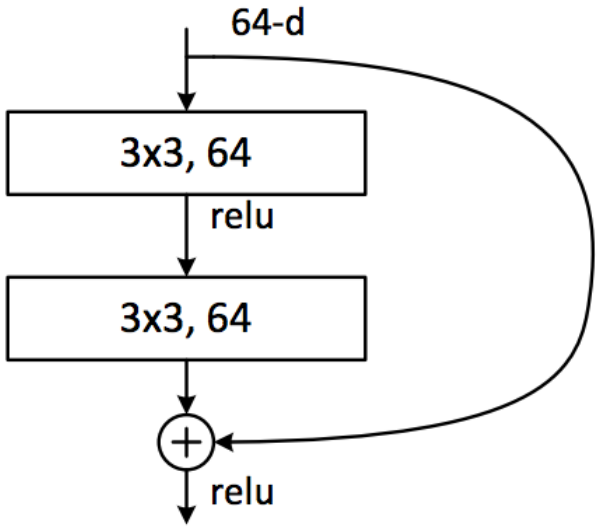
ResNets (residual networks)

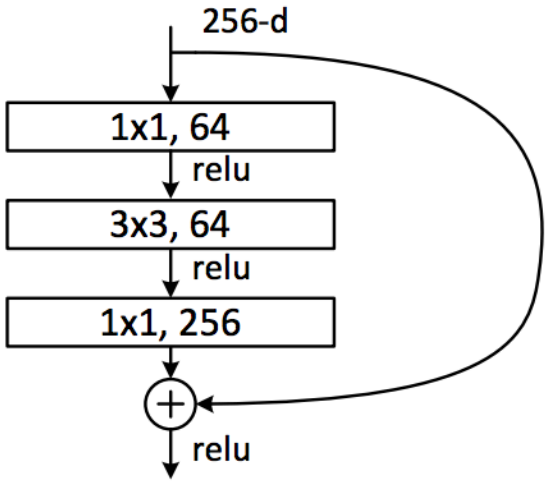


Residual learning: a building block

Deep Residual Learning for Image Recognition
Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, 2015

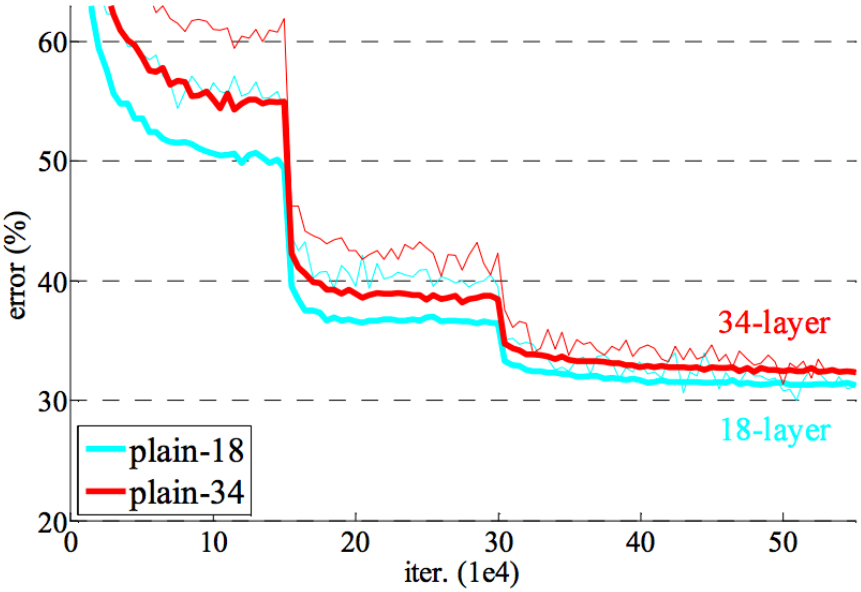
ResNets (residual networks)

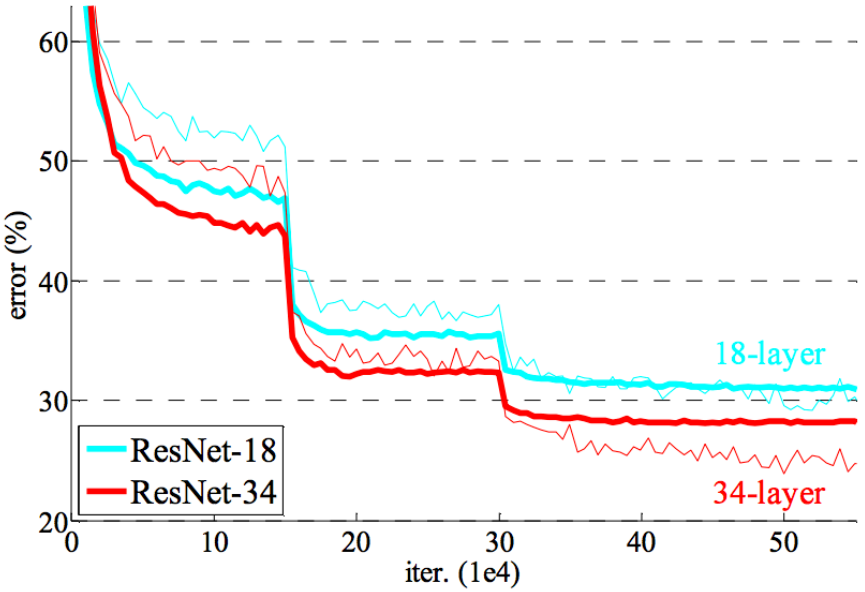




Deep Residual Learning for Image Recognition
Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, 2015

ResNets (residual networks)

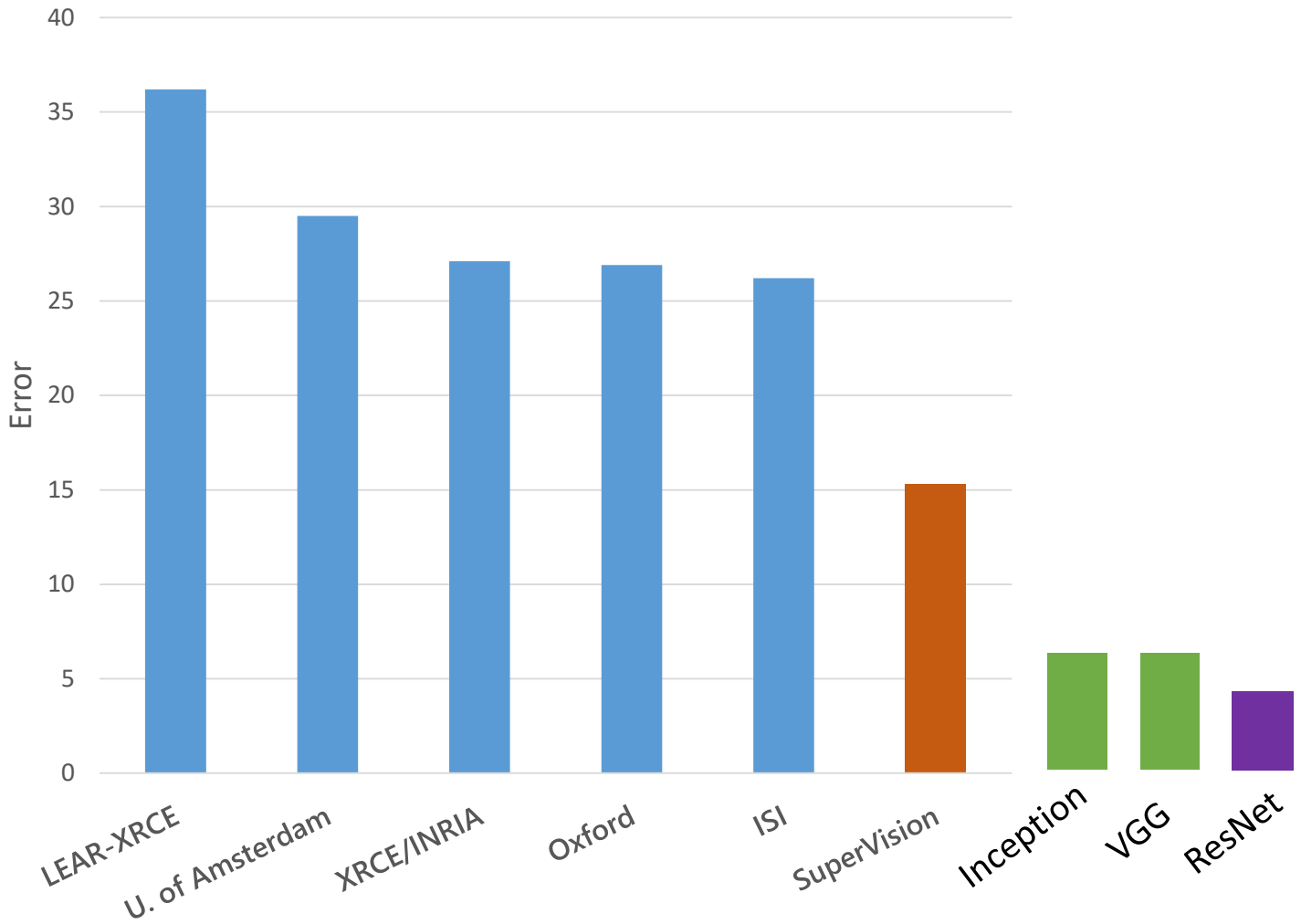




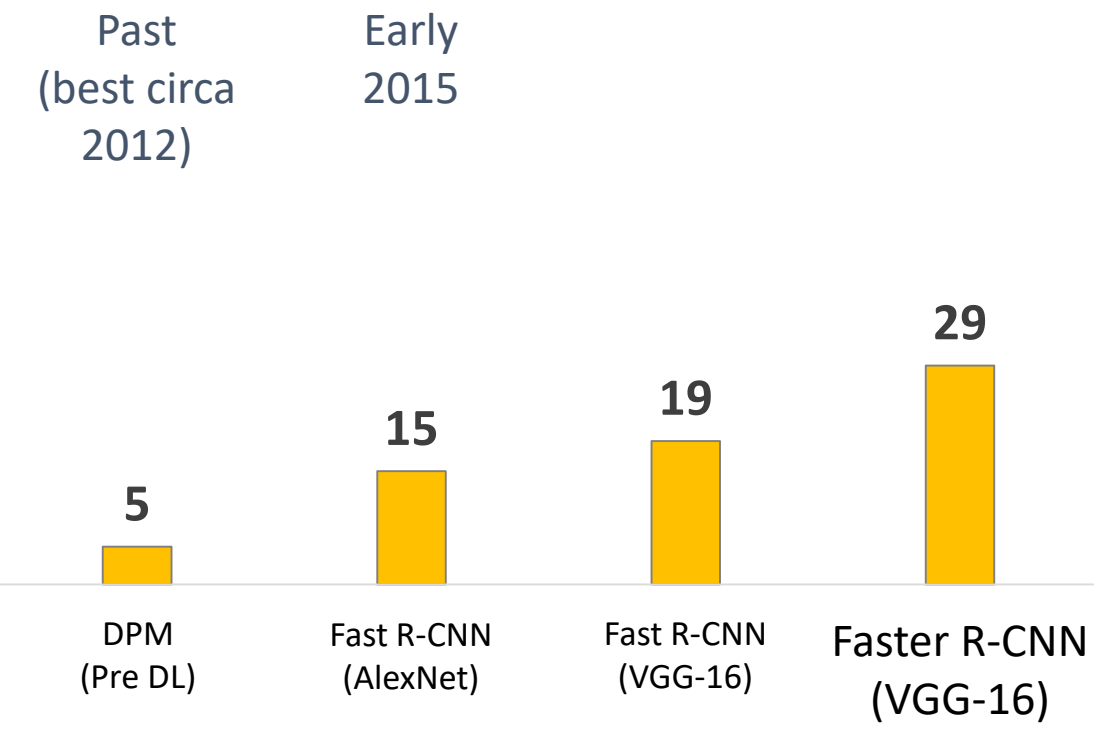
Deep Residual Learning for Image Recognition

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, 2015

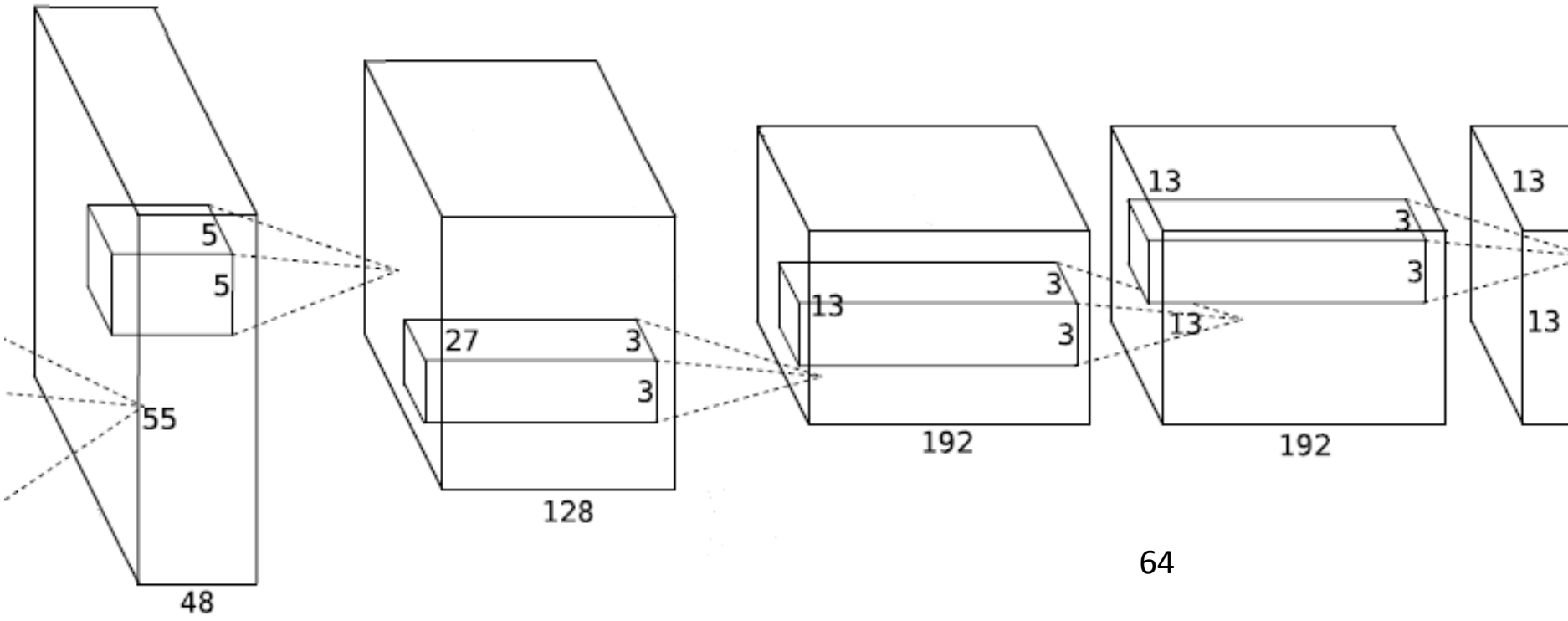
ImageNet 1K


40
35
30
25
20
15
10
5
0
Error
LEAR-XRCE
U. of Amsterdam
XRCE/INRIA
Oxford
ISI
SuperVision
Inception
VGG
ResNet

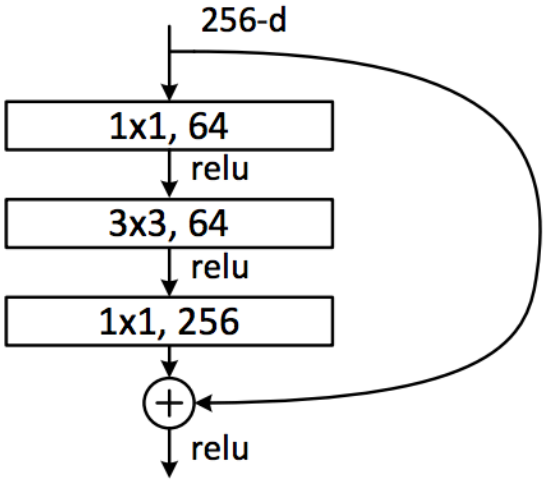
COCO Object Detection Average Precision (%)

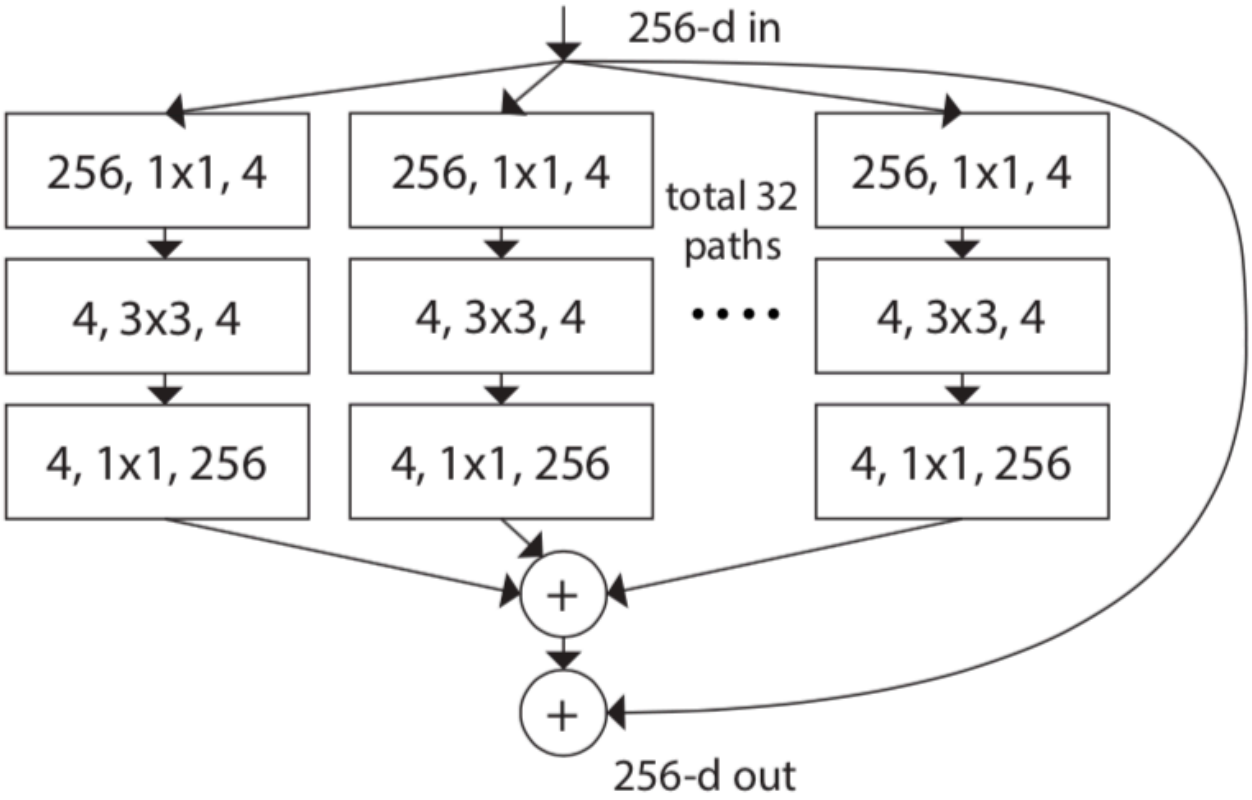


What happens if you want to increase the number of filters?



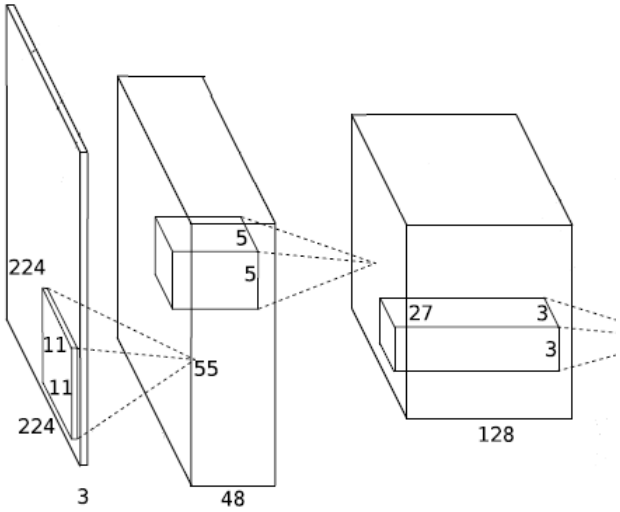
What happens if you want to increase the number of filters?

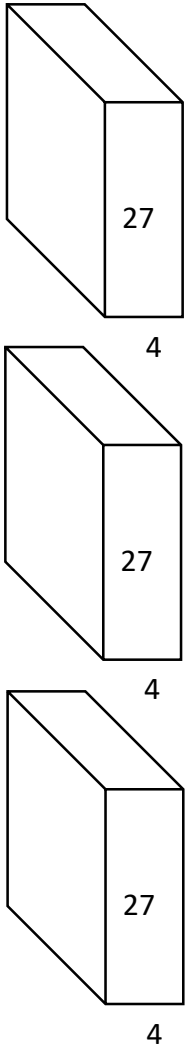




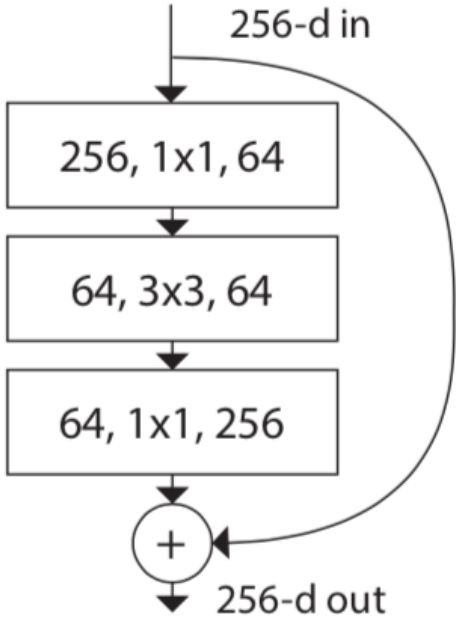
Aggregated residual transformations for deep neural networks
S Xie, R Girshick, P Dollár, Z Tu, K He

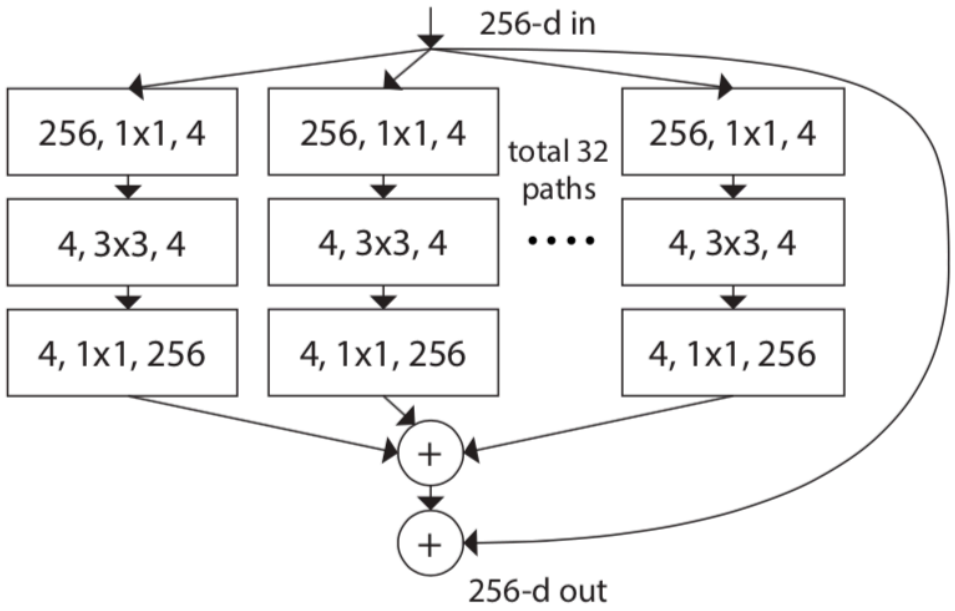
ResNeXt



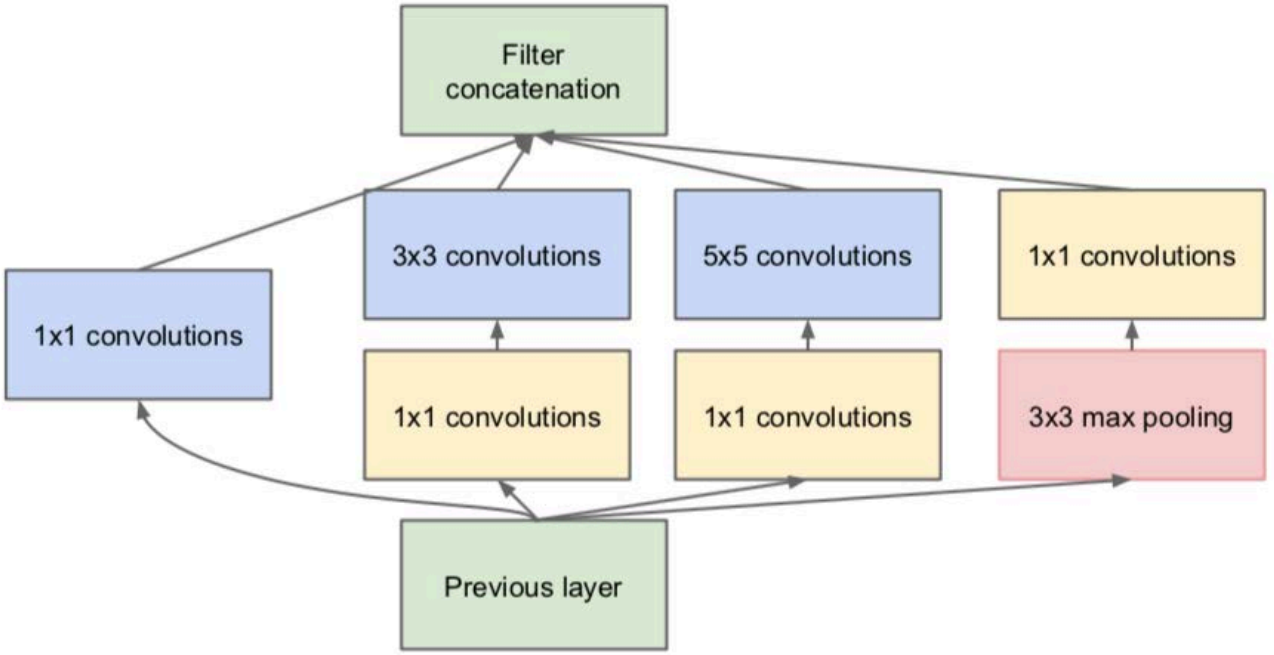


Roughly same complexity!





Remember those Inception modules?



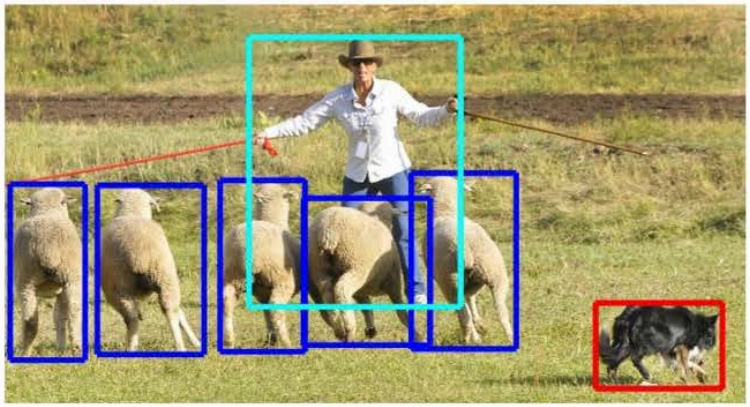
(b) Inception module with dimension reductions

Aggregated residual transformations for deep neural networks
S Xie, R Girshick, P Dollár, Z Tu, K He

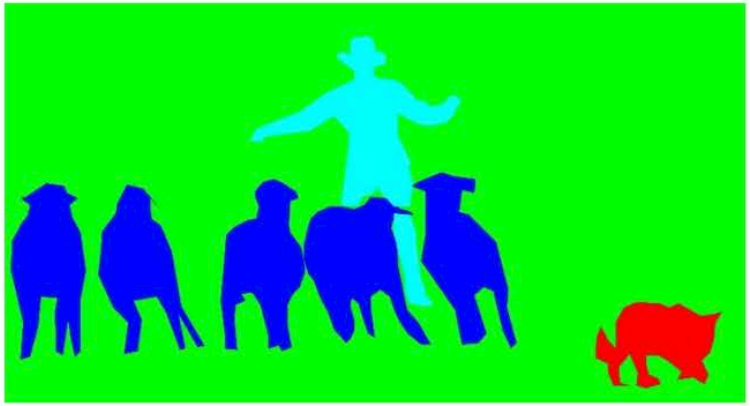
We’ve been talking a lot about detection.

Segmentation

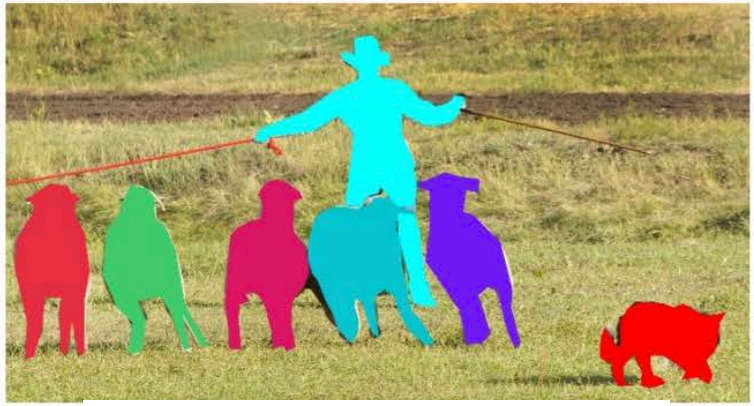
(a) Image classification



(b) Object localization

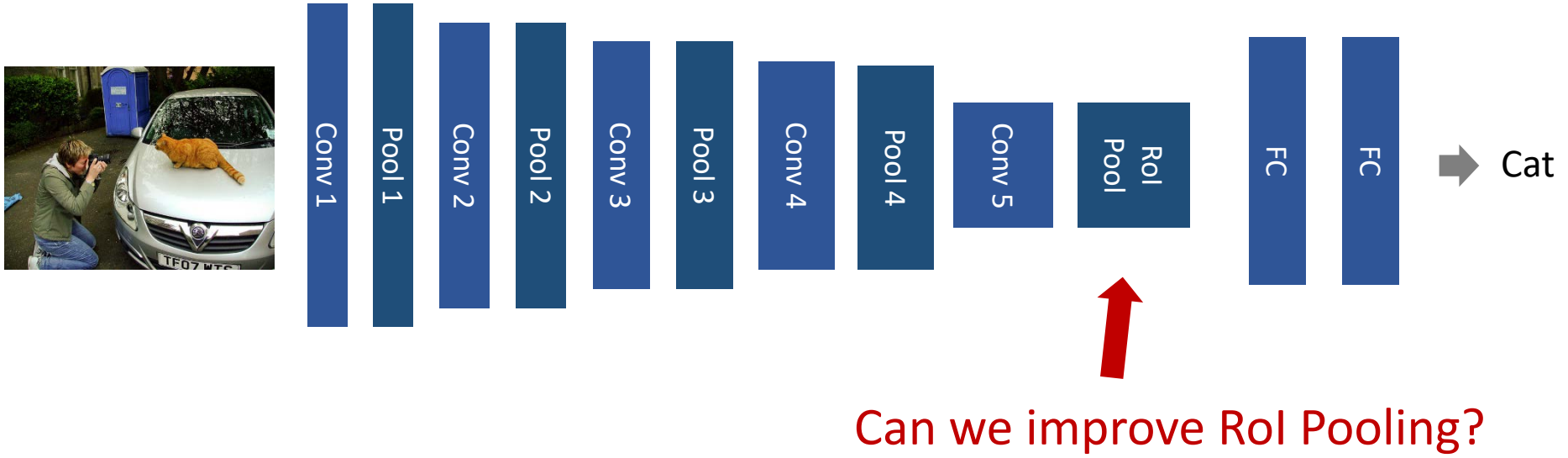


(c) Semantic segmentation

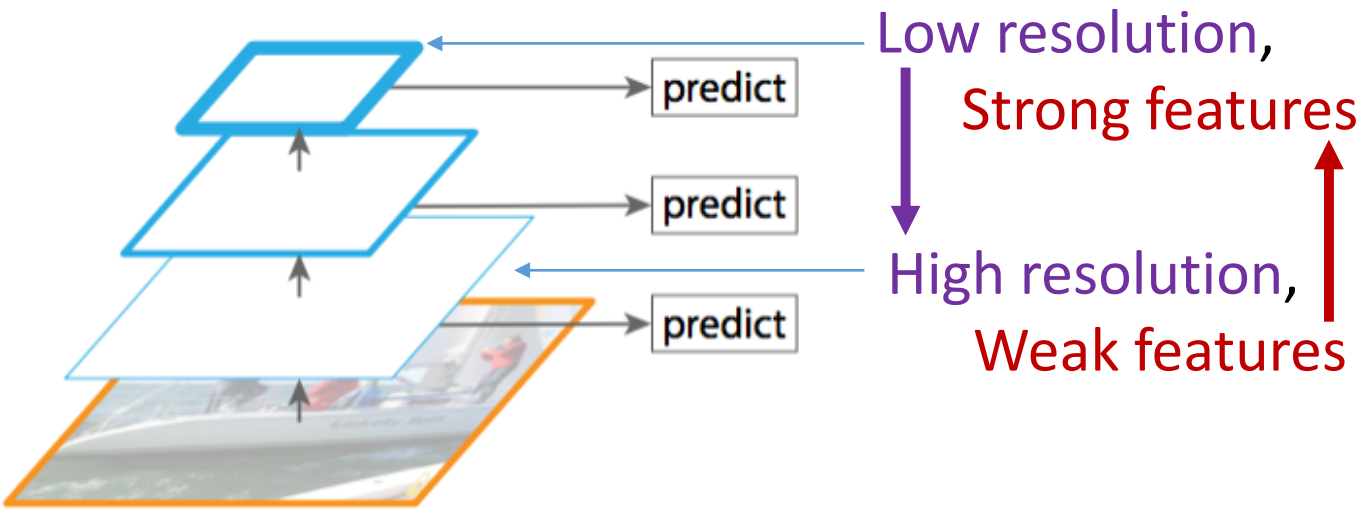


(d) Instance segmentation

Why is segmentation hard with R-CNN type architectures?



Multi-scale

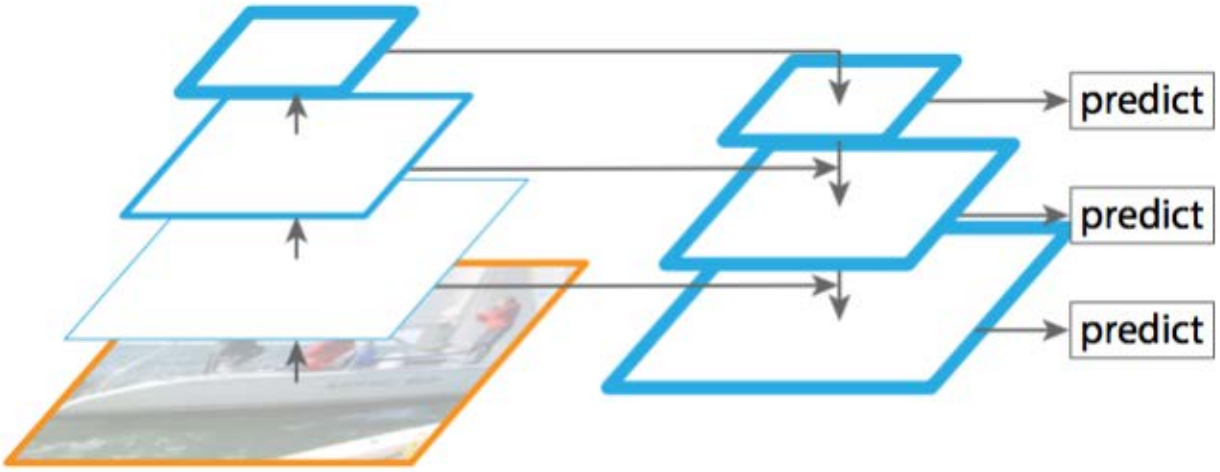


(c) Pyramidal feature hierarchy

Use the internal pyramid – *fast, suboptimal*

(E.g., ≈ SSD, …)

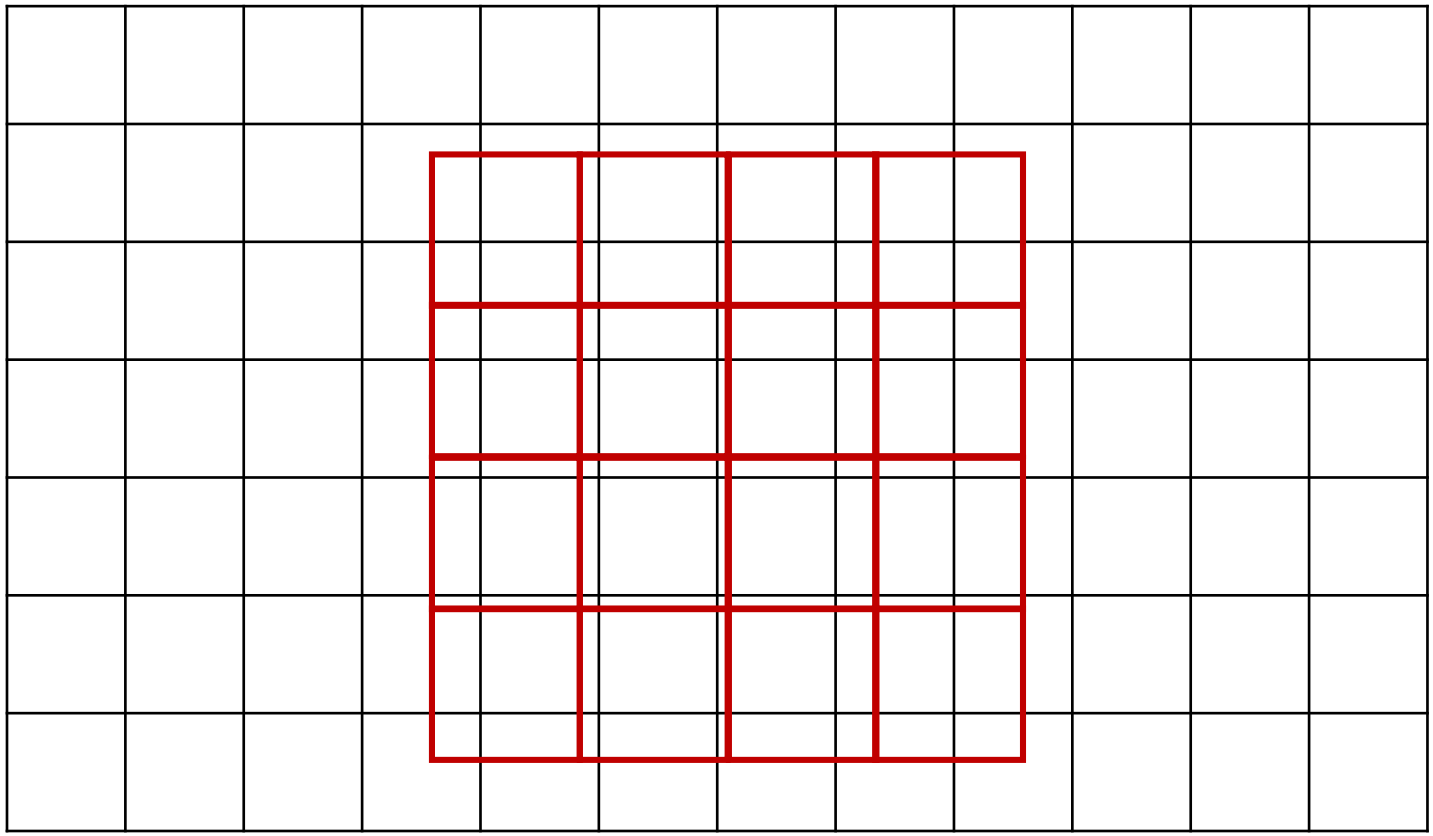
Multi-scale

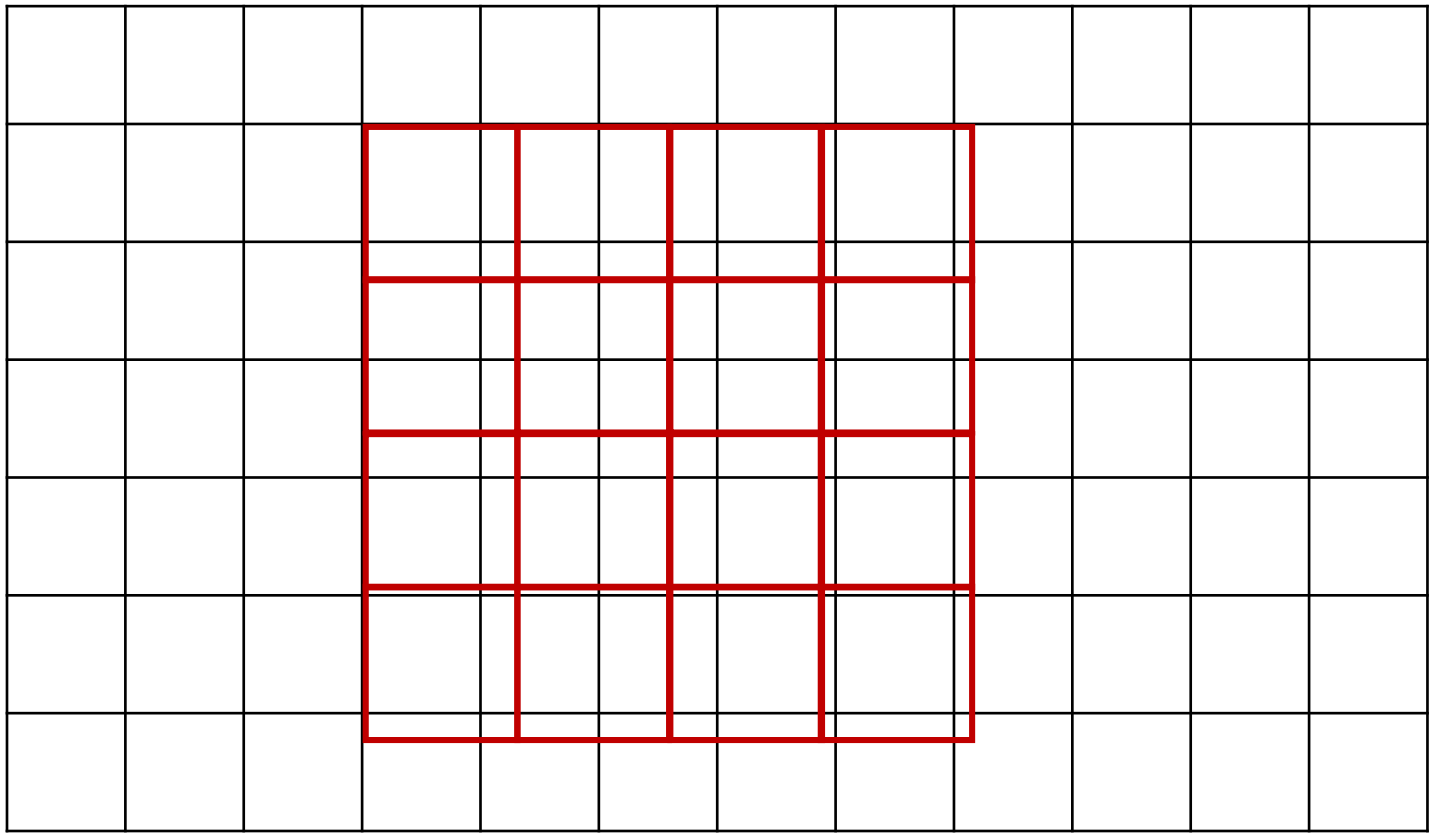


(d) Feature Pyramid Network

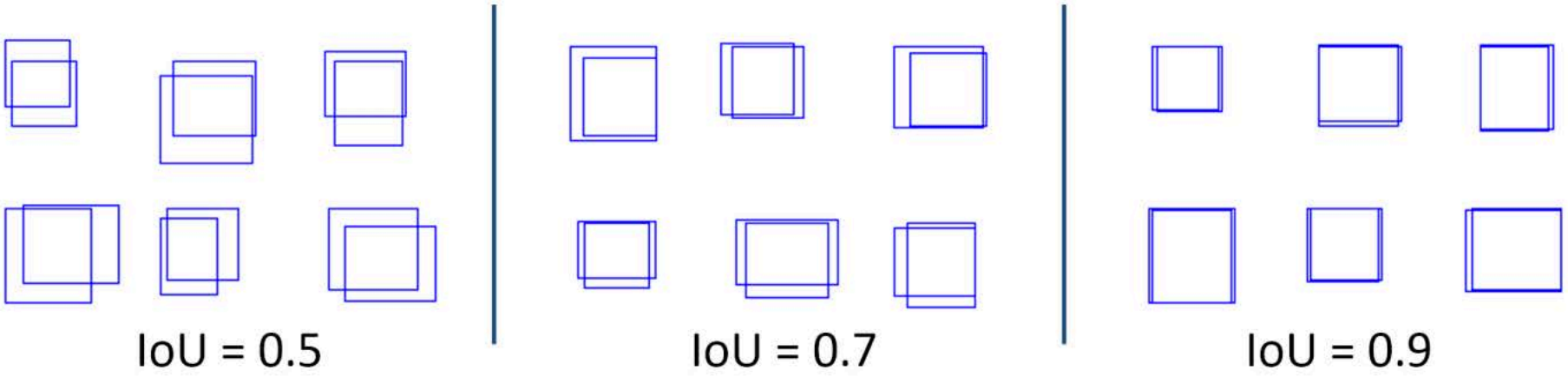
Top-down enrichment of high-res features – *fast, less suboptimal*

Feature Pyramid Networks for Object Detection,
T. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, S. Belongie, CVPR 2017

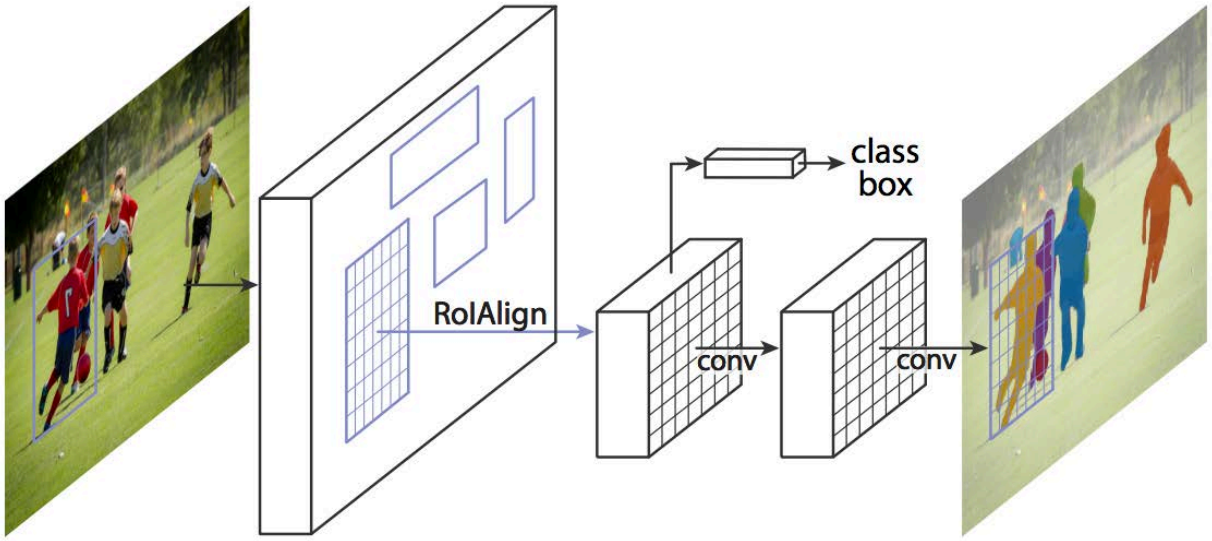




IoU



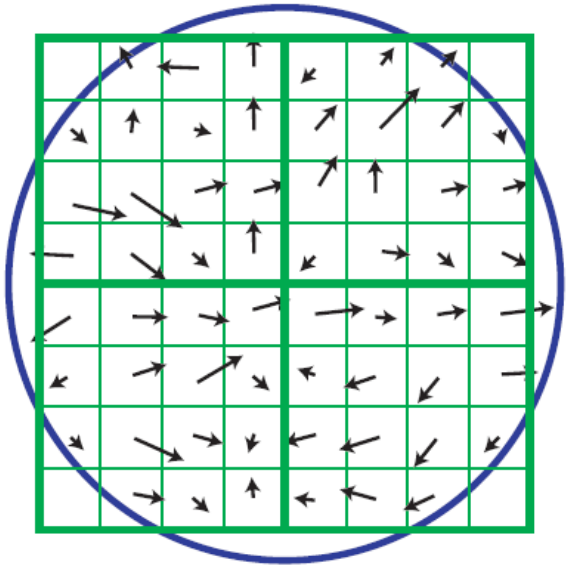
Mask R-CNN



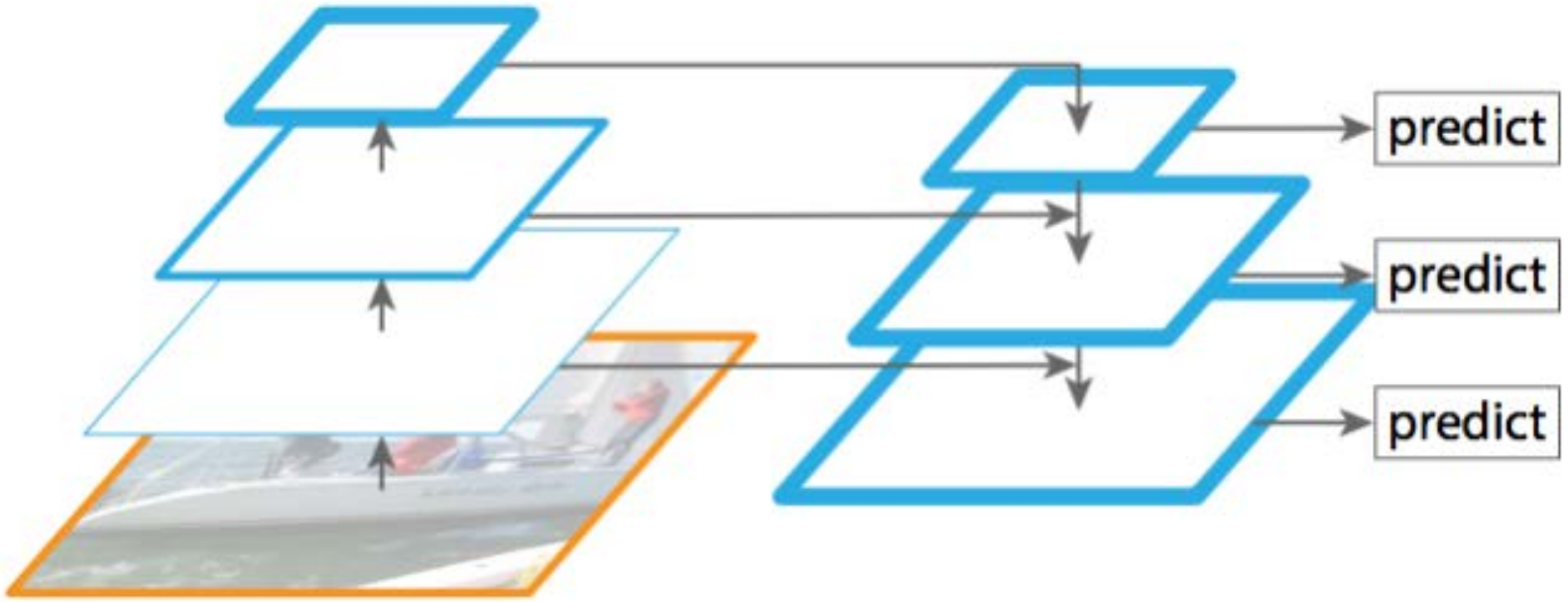
DETECTION,
SEGMENTATION,
KEY POINTS

Mask R-CNN,
Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshick

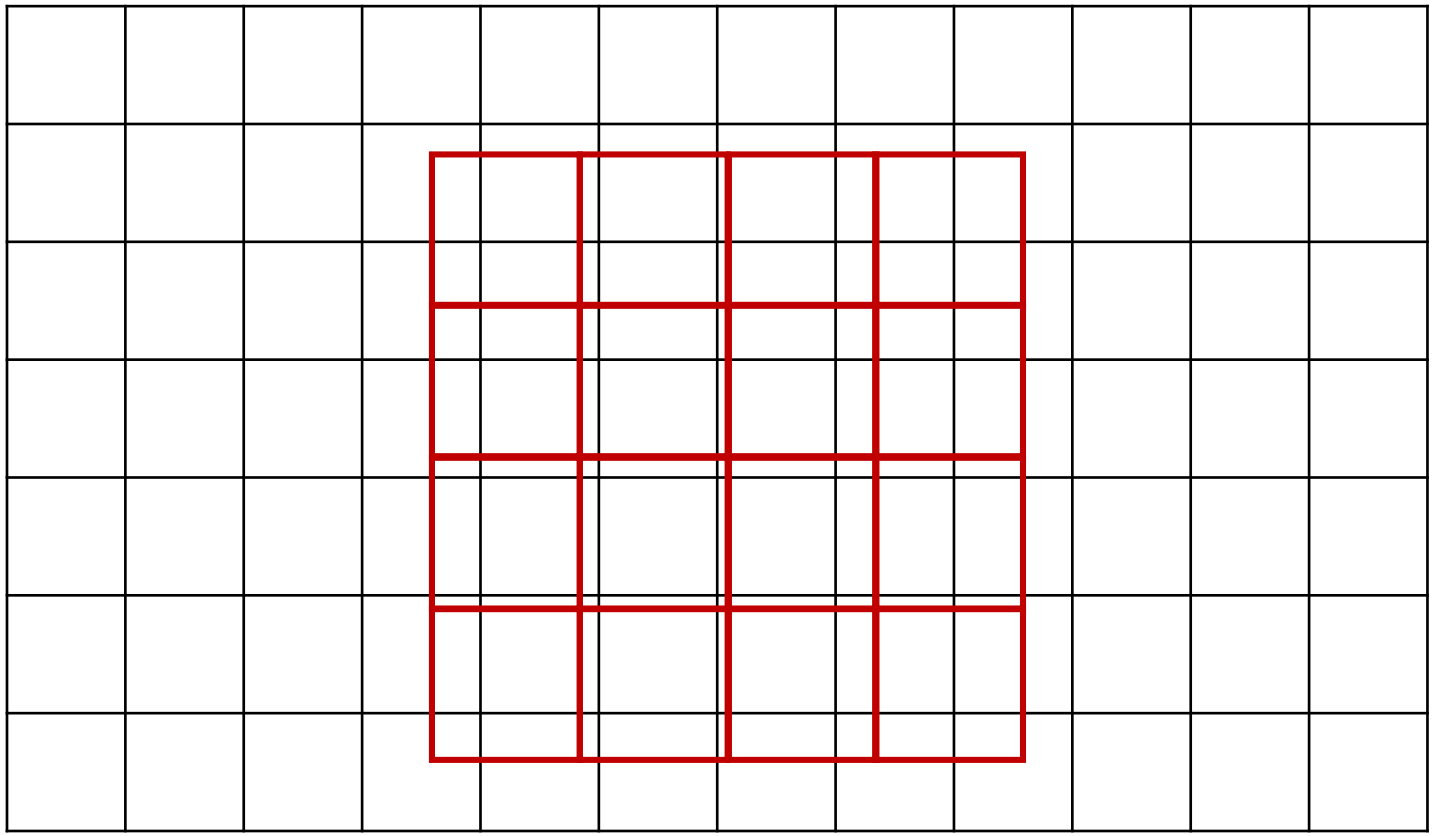
SIFT



RoIAlign layer



Feature Pyramid Network

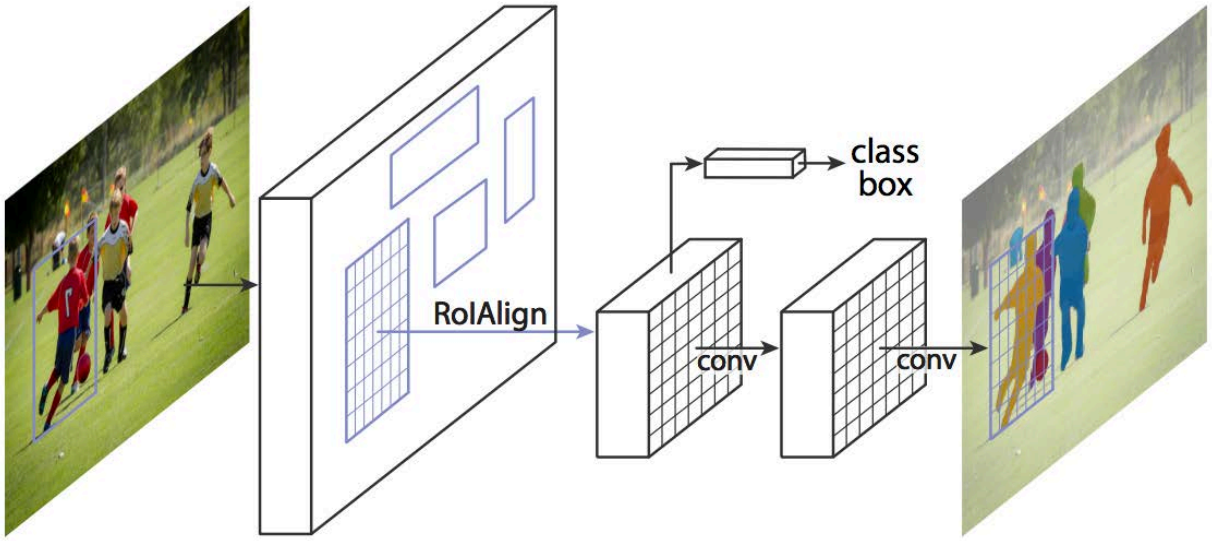


	align?	bilinear?	agg.	AP	AP_{50}	AP_{75}
RoIPool [12]			max	26.9	48.8	26.4
RoIWarp [10]		✓	max	27.2	49.2	27.1
		✓	ave	27.1	48.9	27.1
RoIAlign	✓	✓	max	**30.2**	**51.0**	**31.8**
	✓	✓	ave	**30.3**	**51.2**	**31.5**

Mask results with various RoI layers

Mask R-CNN

Separate classification and mask prediction



Mask R-CNN, Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshick

Finding object candidates

Use low-level cues…

Interest points all over again?

Segmentation As Selective Search for Object Recognition,
van de Sande, Uijlings, Gevers, Smeulders, ICCV 2011

Why region proposals?

Why not sliding window?

“One stage” detectors

Focal Loss for Dense Object Detection,
Tsung-Yi Lin Priya Goyal Ross Girshick Kaiming He, Piotr Dollar, ICCV 2017

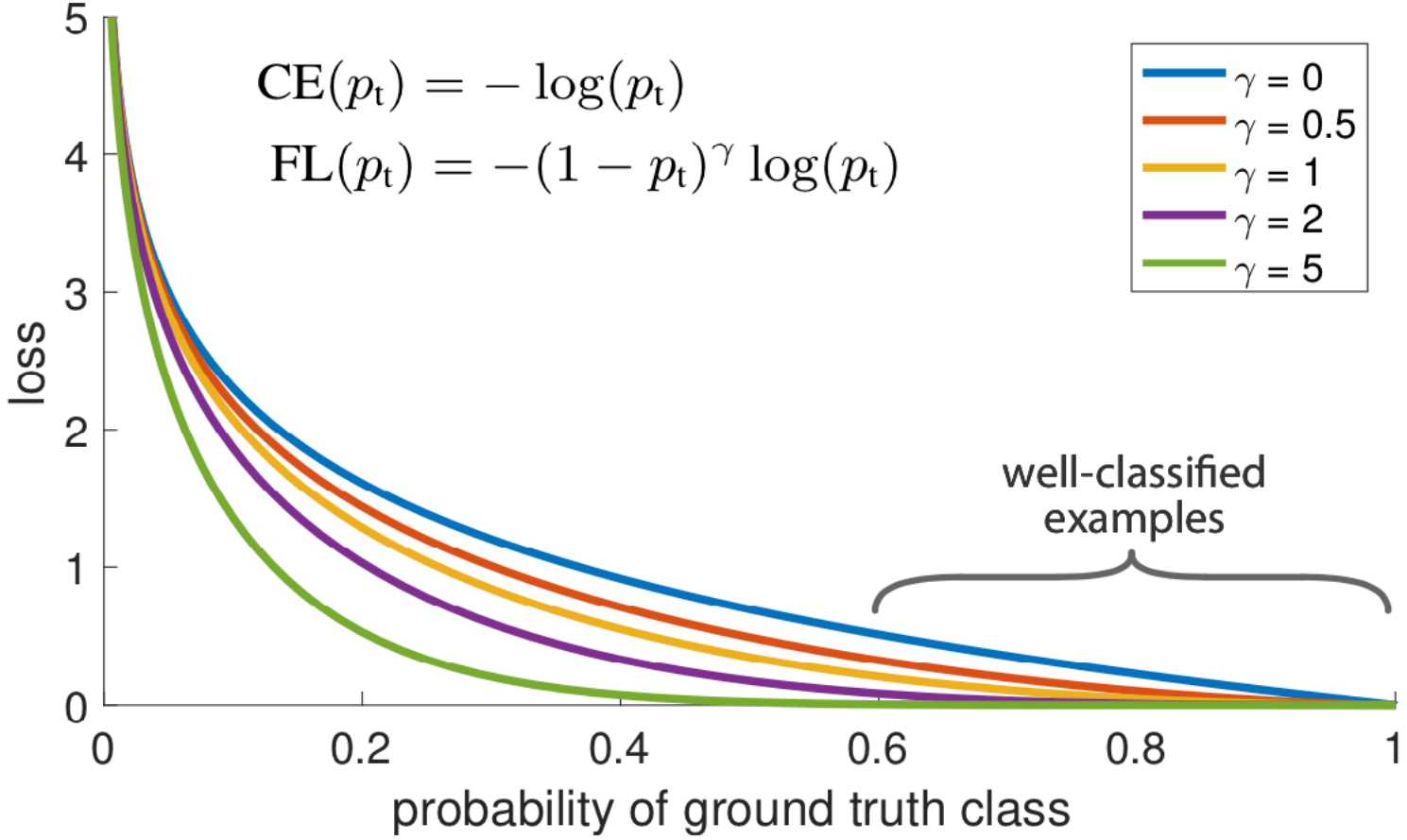
“One stage” detectors

Big class imbalance:

1 positive for ~1-10k negatives

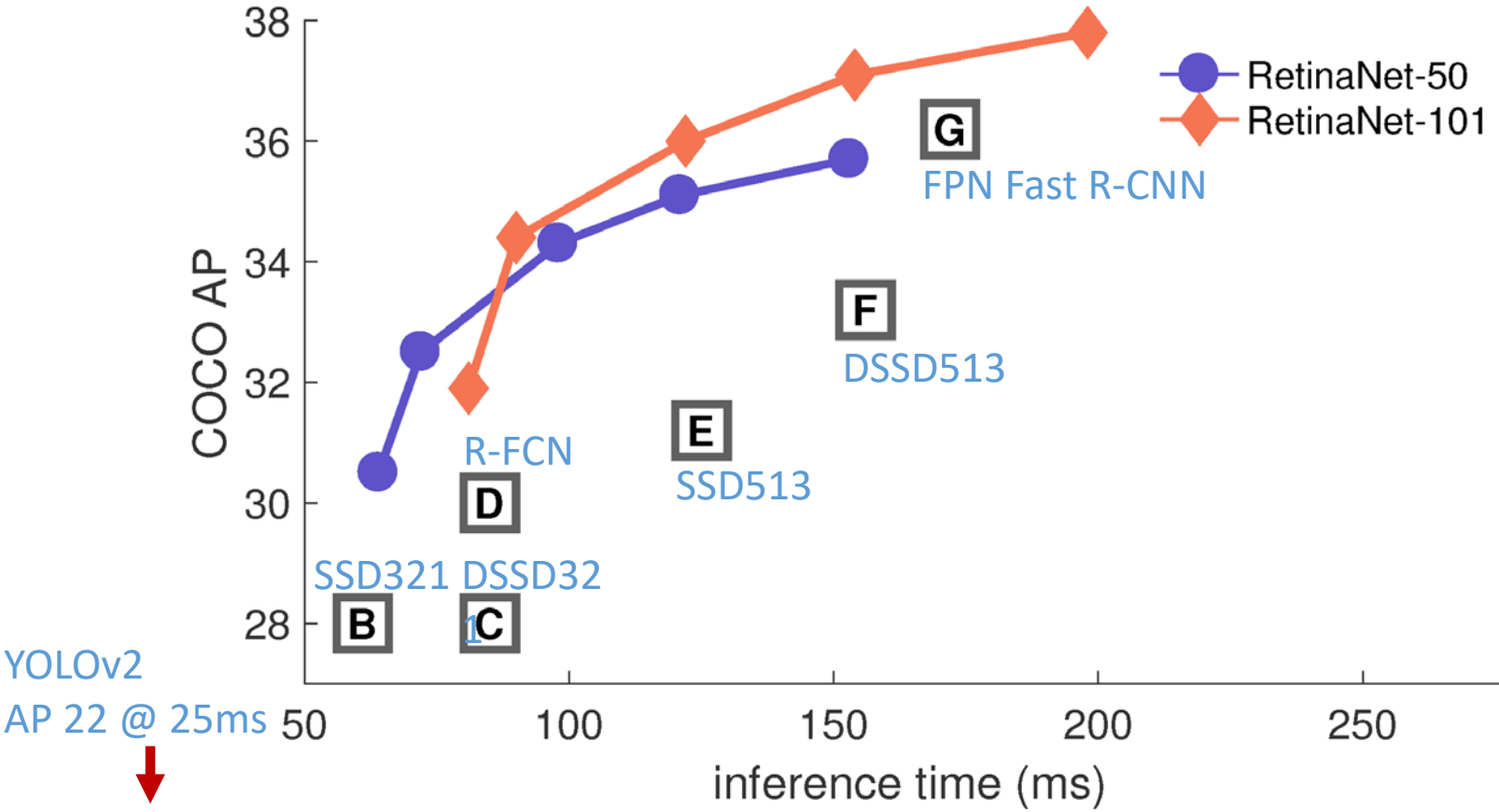
Focal Loss for Dense Object Detection,
Tsung-Yi Lin Priya Goyal Ross Girshick Kaiming He, Piotr Dollar, ICCV 2017

“One stage” detectors

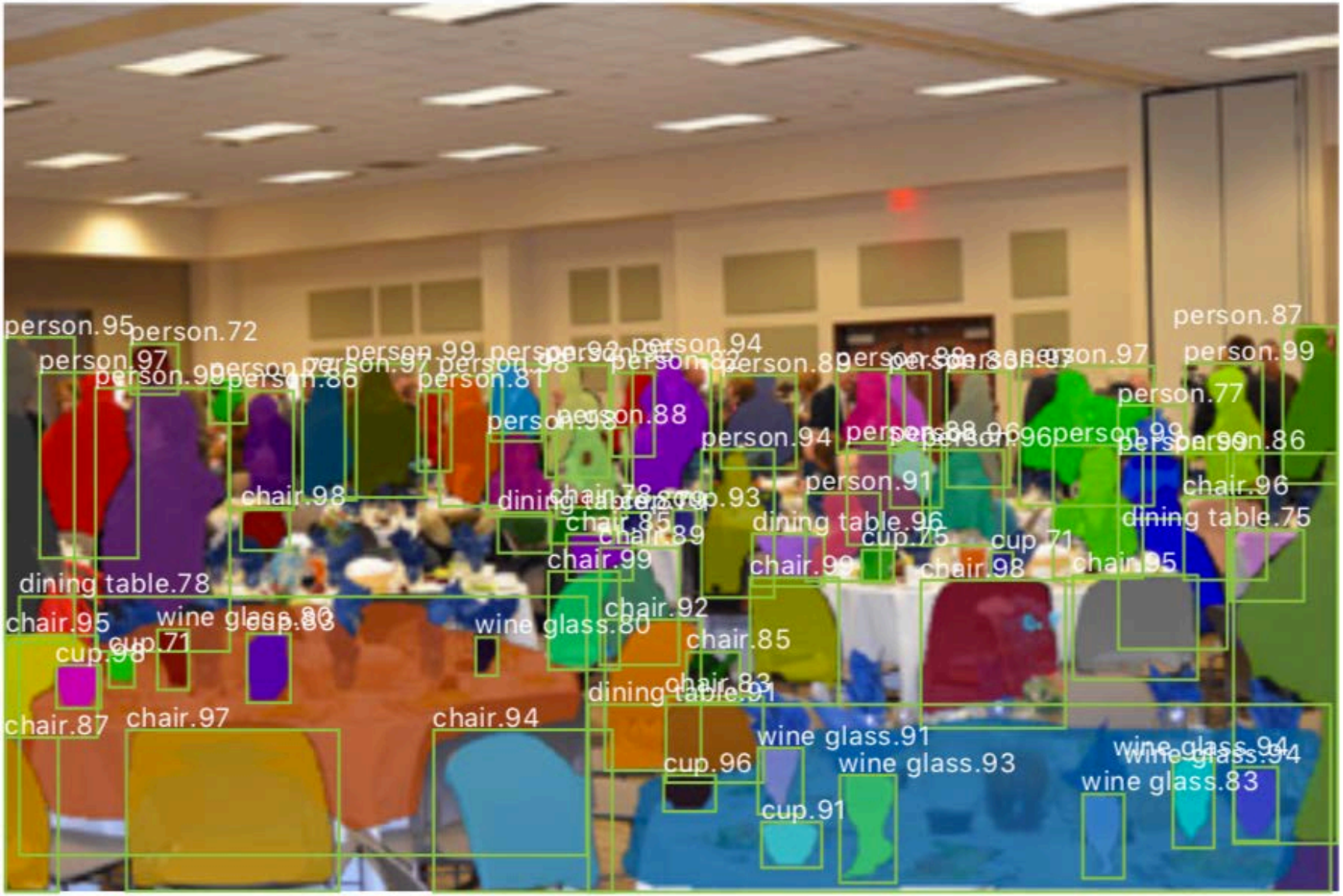


Focal Loss for Dense Object Detection,
Tsung-Yi Lin Priya Goyal Ross Girshick Kaiming He, Piotr Dollar, ICCV 2017

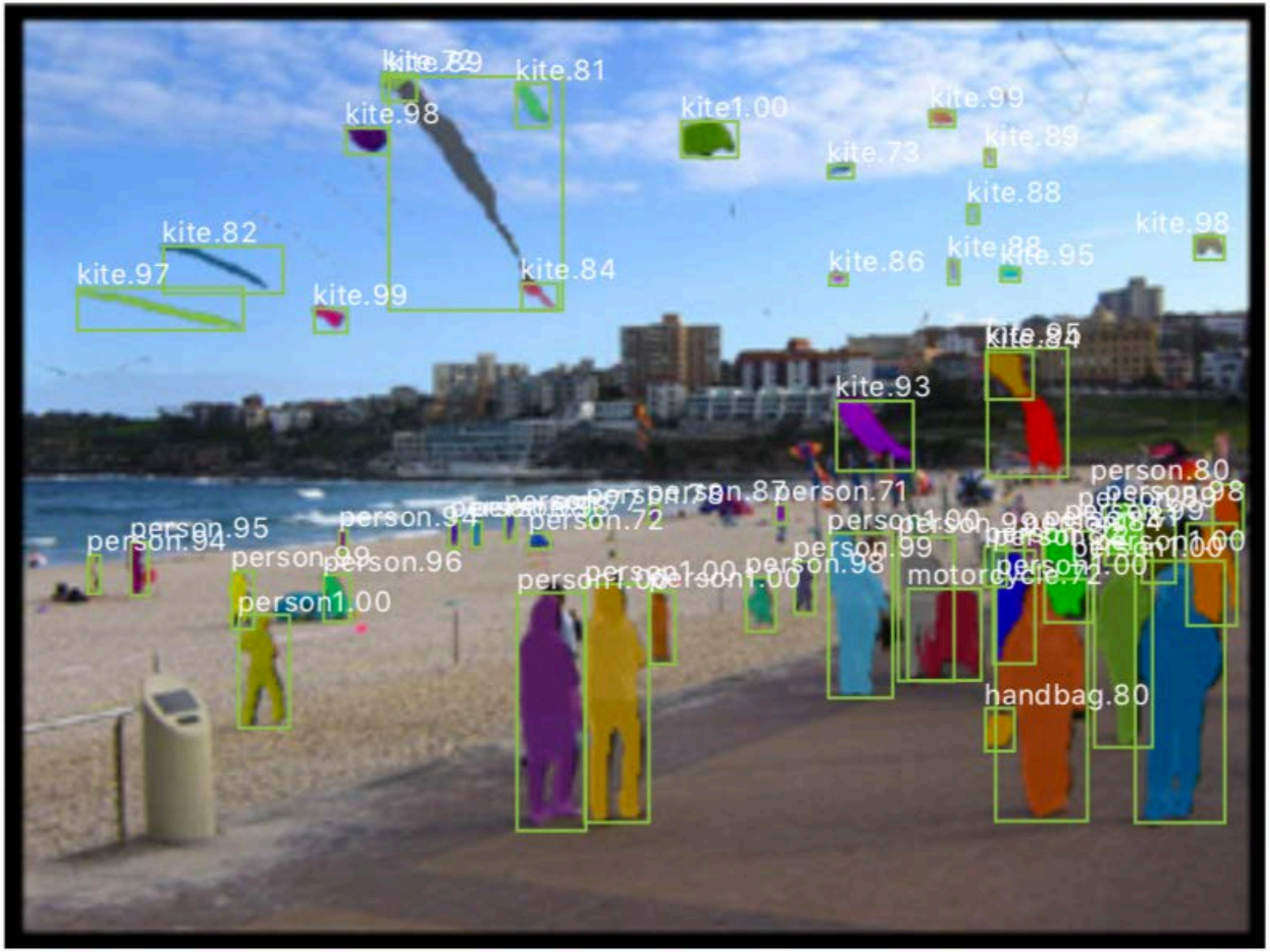
Speed/Accuracy Tradeoff



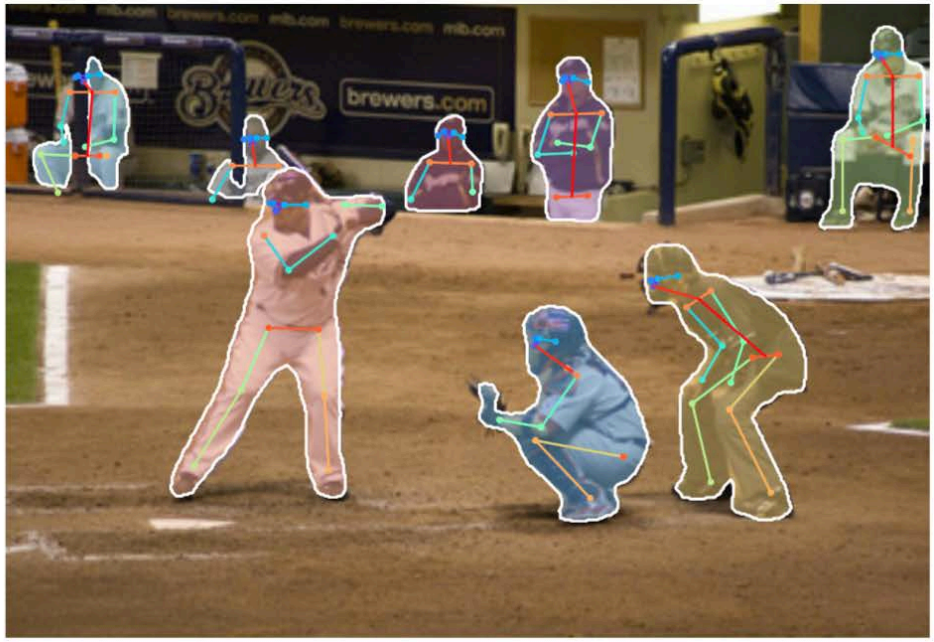
Where are we now?



Mask R-CNN,
Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshick



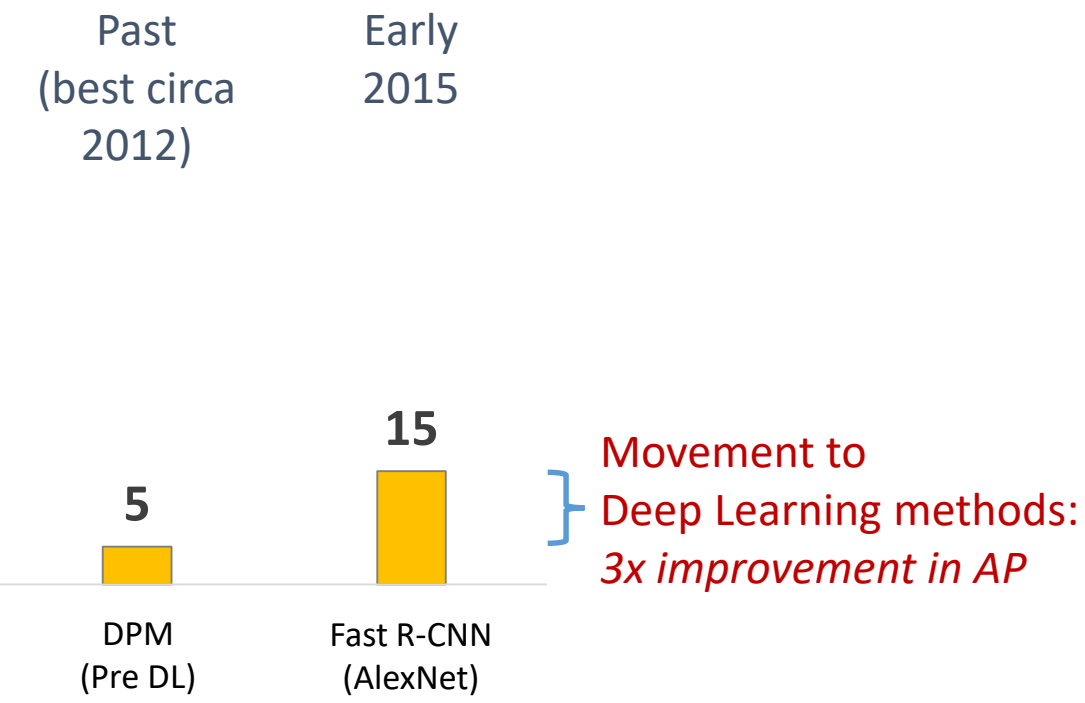
Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshick



Mask R-CNN,
Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshick

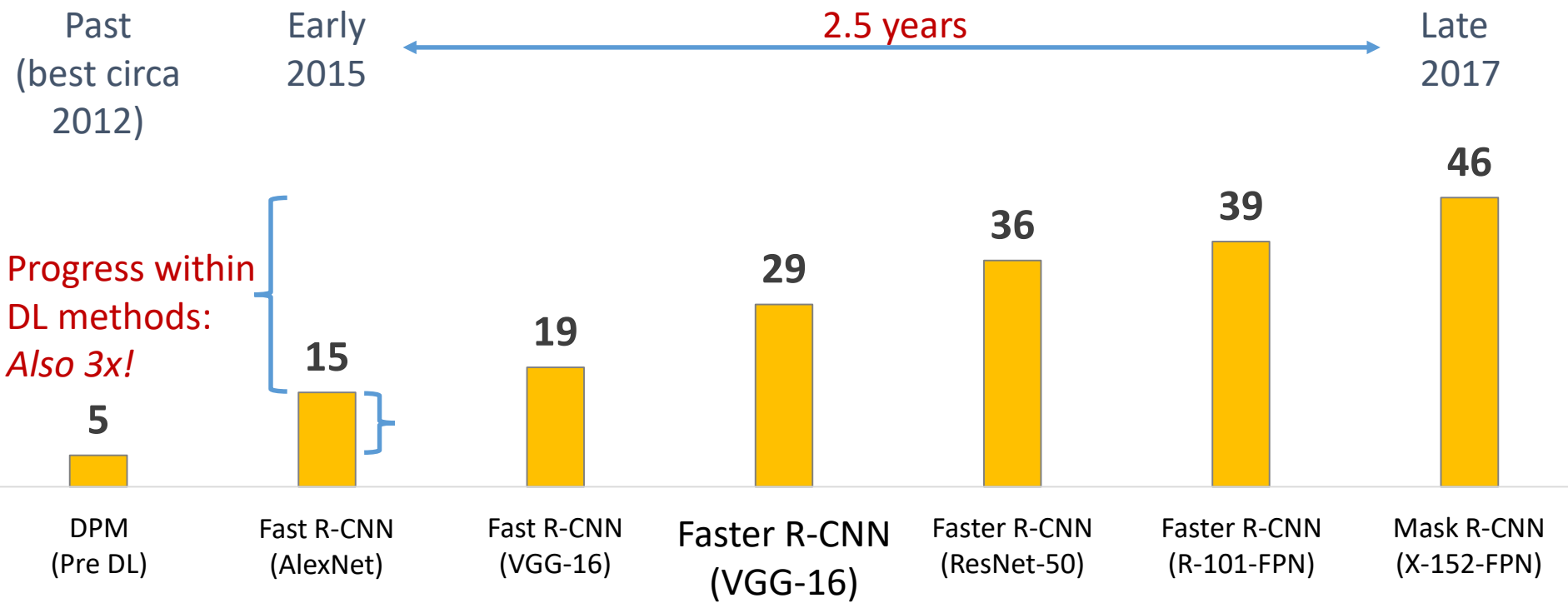

Scrabble
Scrabble

COCO Object Detection Average Precision (%)



Slides by Ross Girshick

COCO Object Detection Average Precision (%)



Slides by Ross Girshick

Lessons

- Your algorithm doesn’t need to be perfect
  - Others will fix it later
- Simple beats complex EVERY time
  - The best papers have simple ideas with big impact
- Read
  - Then read more

Imagenet Classification with Deep Convolutional Neural Networks,
Krizhevsky, Sutskever, and Hinton, *NIPS* 2012

Very deep convolutional networks for large-scale image recognition,
Karen Simonyan, Andrew Zisserman, *ICLR* 2015

Going deeper with convolutions,
Szegedy, Liu, Jia, Sermanet, Reed, Anguelov, Erhan, Vanhoucke, Rabinovich, *CVPR* 2015

Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition
Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, *ECCV* 2014

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

Fast R-CNN
Ross Girshick, *ICCV* 2015

Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks
S Ren, K He, R Girshick, J Sun, *NIPS* 2015

Deep Residual Learning for Image Recognition
Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, *CVPR* 2016

Aggregated residual transformations for deep neural networks
S Xie, R Girshick, P Dollár, Z Tu, K He, *CVPR* 2017

Feature Pyramid Networks for Object Detection,
T. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, S. Belongie, *CVPR* 2017

Mask R-CNN,
Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshick, *ICCV* 2017

Focal Loss for Dense Object Detection,
Tsung-Yi Lin Priya Goyal Ross Girshick Kaiming He, Piotr Dollar, *ICCV* 2017

Looking forward…

New evaluations

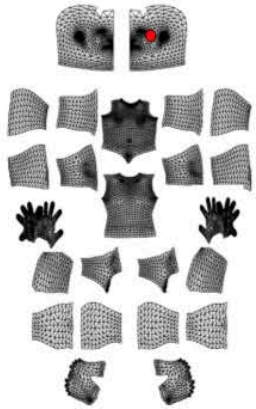
(a) image (b) semantic segmentation

(c) instance segmentation (d) panoptic segmentation

Panoptic Segmentation,
Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, Piotr Dollar

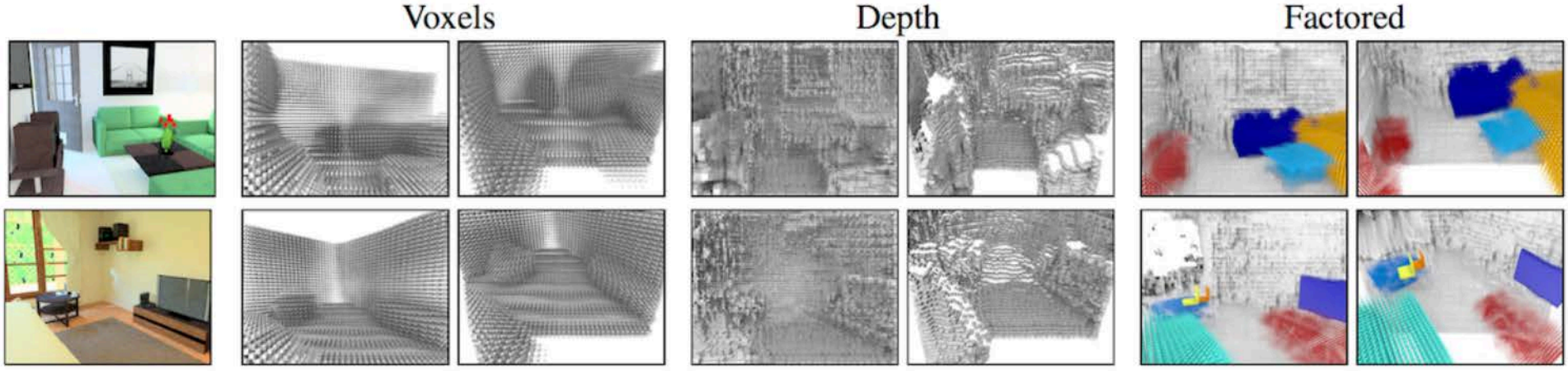
3D

DensePose-RCNN Results



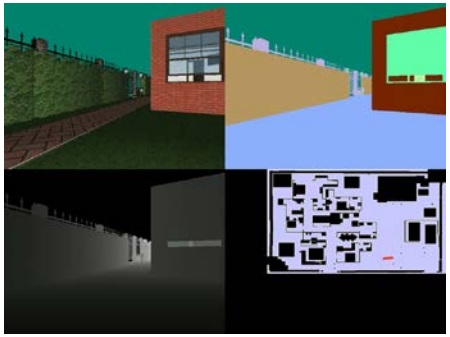
DensePose COCO Dataset

DensePose: Dense Human Pose Estimation In The Wild, Rıza Alp Guler, Natalia Neverova, Iasonas Kokkinos, 2018



Factoring Shape, Pose, and Layout From the 2D Image of a 3D Scene
Shubham Tulsiani, Saurabh Gupta, David Fouhey, Alexei Efros, Jitendra Malik, 2017

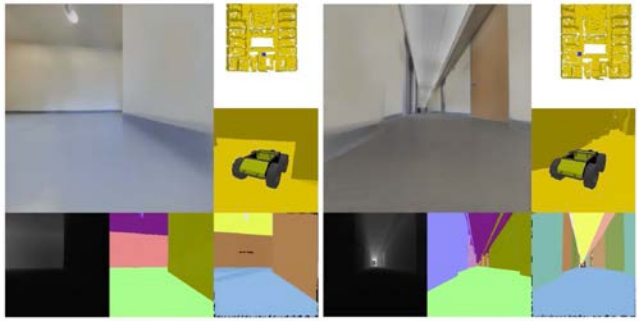
Environments



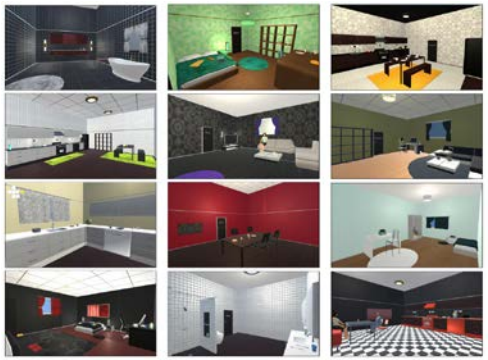
House3D
(Wu et al., 2017)

AI2-THOR
(Kolve et al., 2017)

MINOS
(Savva et al., 2017)



Gibson
(Zamir et al., 2018)

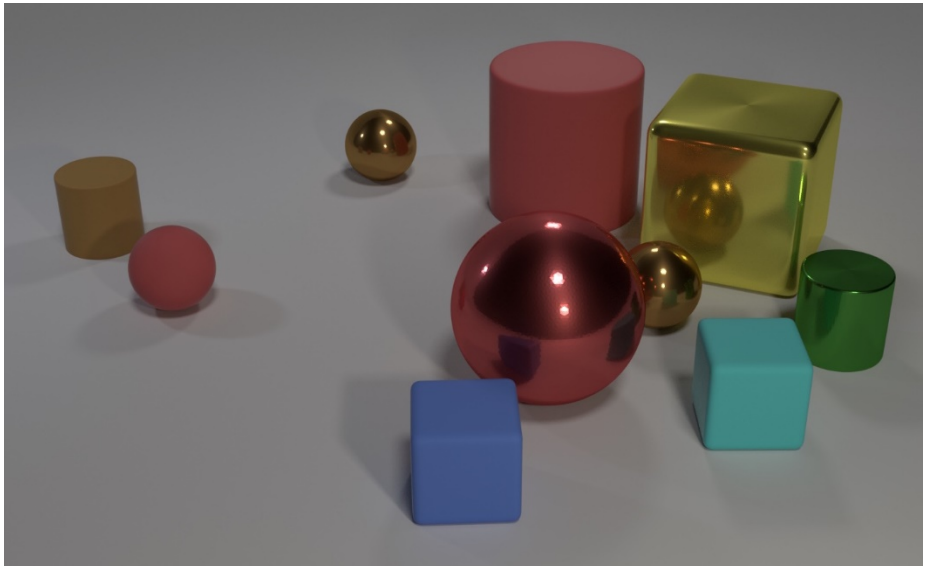


CHALET
(Yan et al., 2018)

Reasoning

Clever Hans, 1907

Reasoning



Q: Are there an **equal number** of **large** things and **metal spheres**?

Q: **What size** is the **cylinder that is left of** the **brown metal** thing **that is left of** the **big sphere**?

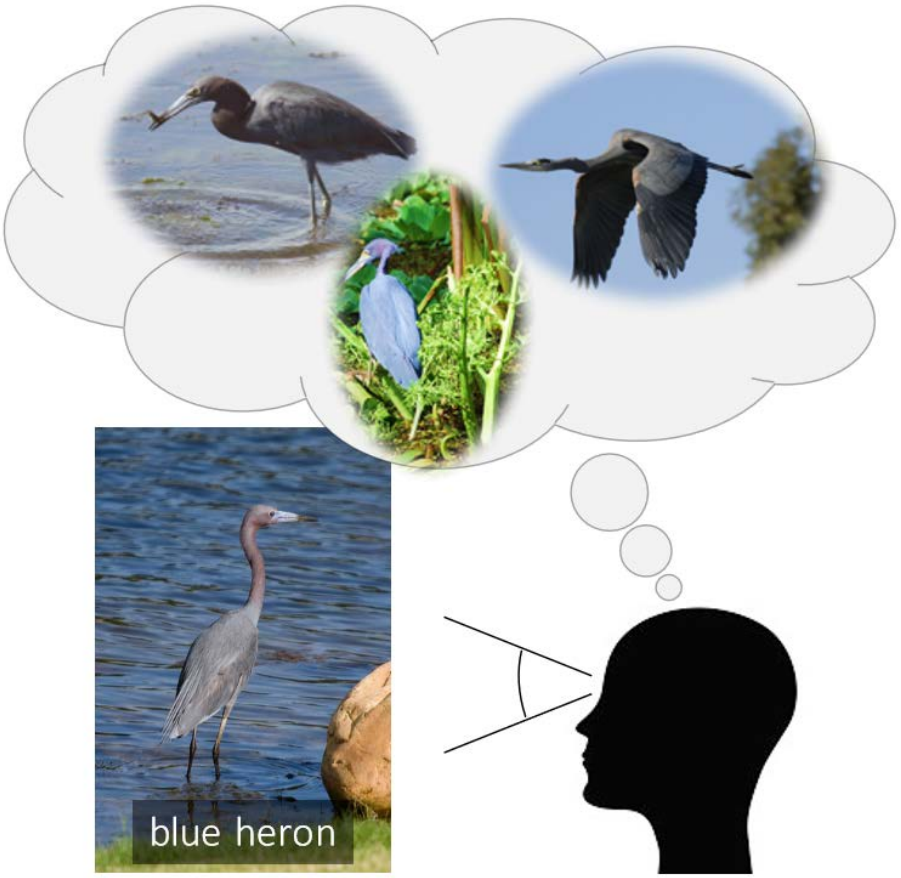
Q: There is a **sphere** with the **same size as** the **metal cube**; is it **made of the same material** as the **small red sphere**?

Q: **How many** objects are **either small cylinders or metal things**?

Attributes **Counting** **Comparison** **Spatial Relationships** **Logical Operations**

CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning, Johnson et al., CVPR 2017

Low-shot Learning



Low-Shot Learning from Imaginary Data, Wang, Girshick,Hebert, Hariharan, 2018

Interpretability

Interpretability

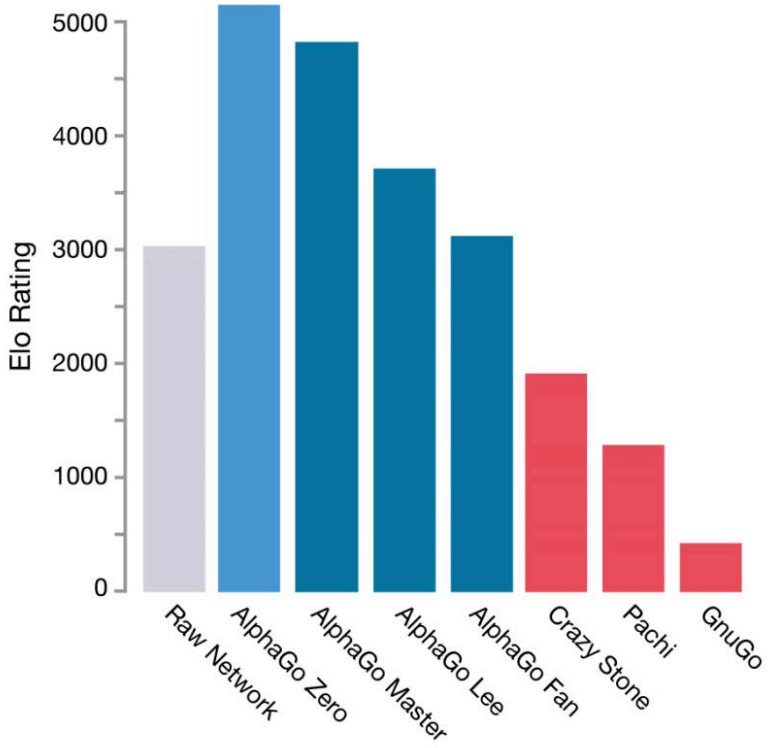
Pierre Stock,
Moustapha Cisse

Onwards and upwards

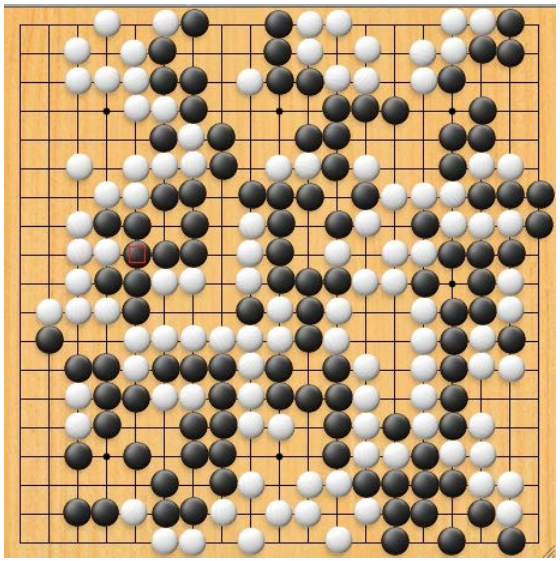
Thanks!

Reasoning?

ELF OpenGo
Yuandong Tian, Jerry Ma*, Qucheng Gong*,
Shubho Sengupta, Zhuoyuan Chen, Larry Zitnick



Mastering the Game of Go without Human Knowledge, Silver et al., 2017

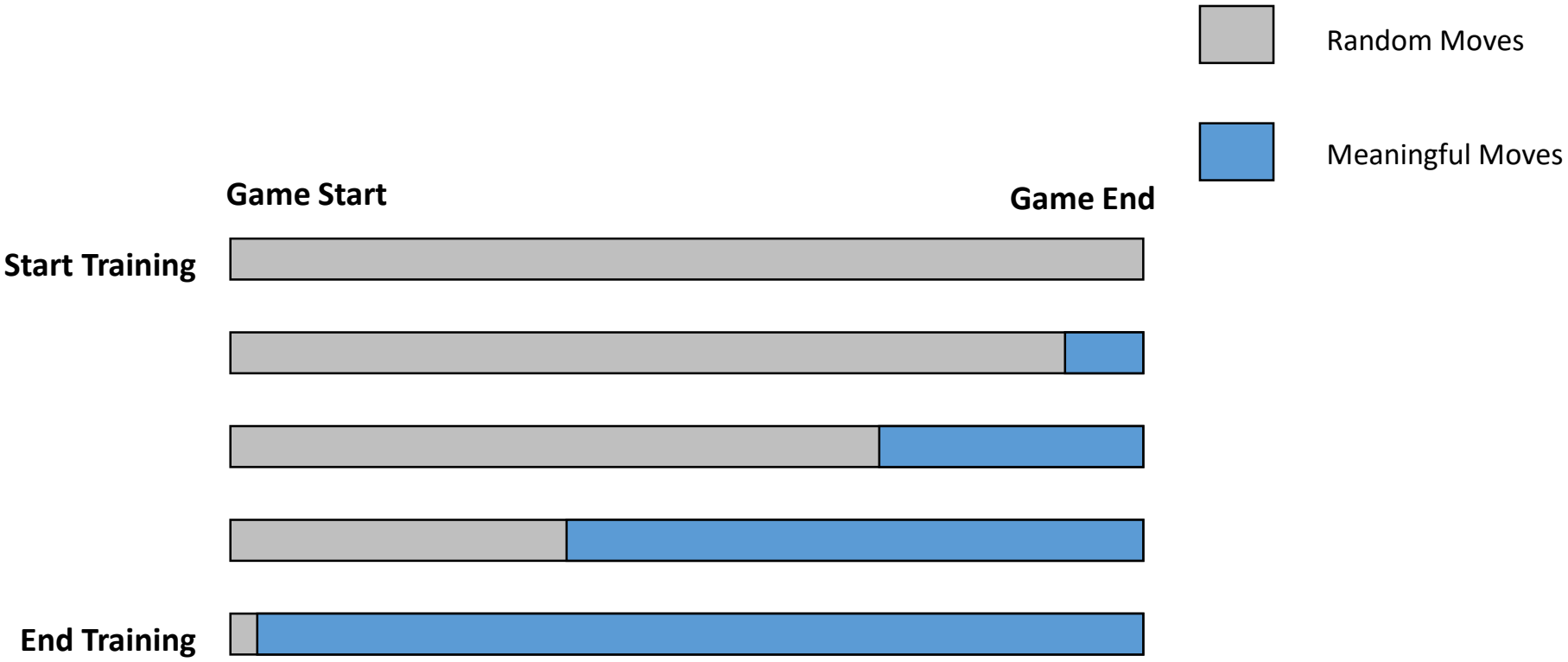


s

$\mathbf{p}_\theta(s)$

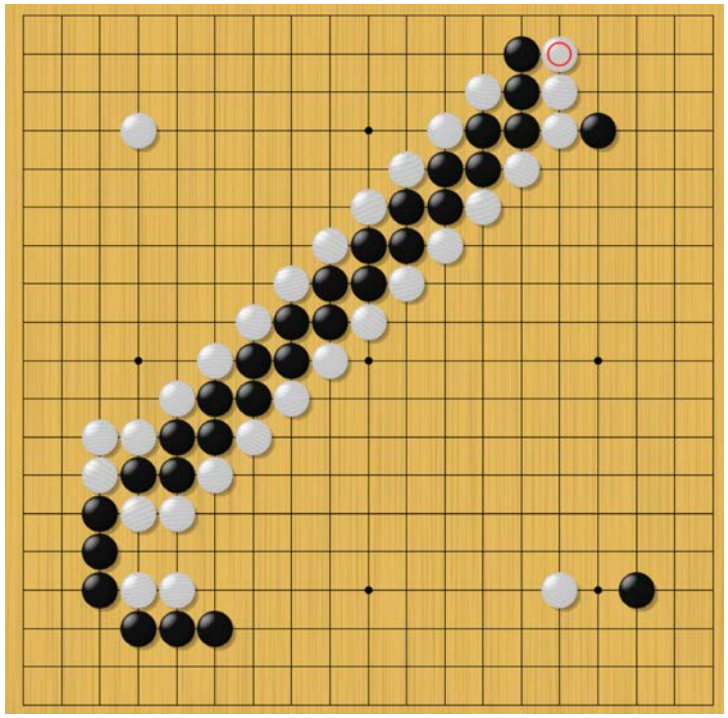
$V_\theta(s)$

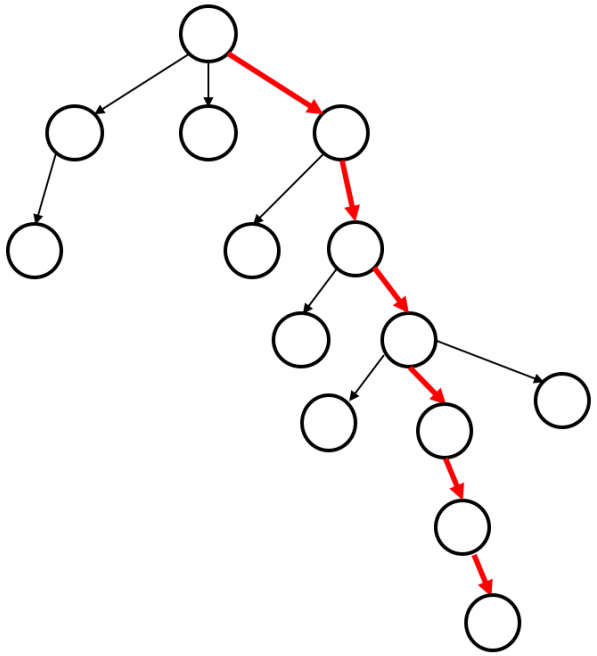
+ MCTS



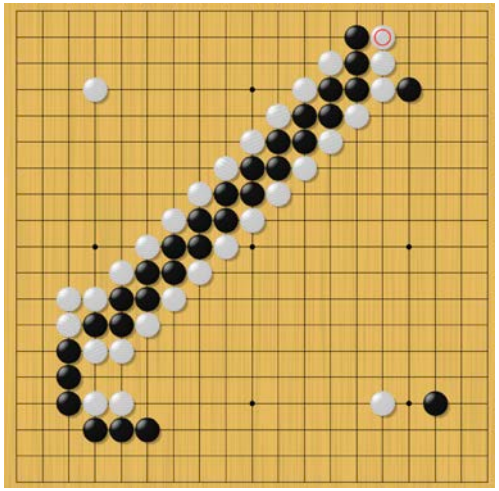
Already dan level even if the opening doesn’t make much sense.

Ladders

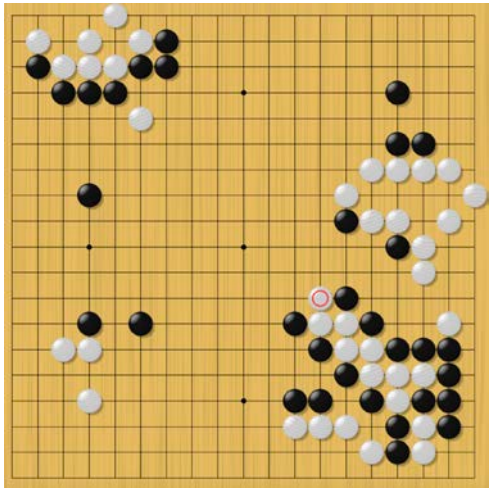




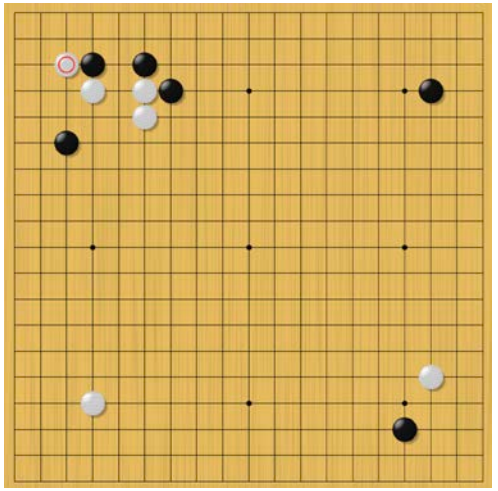
There is only one long path that is correct



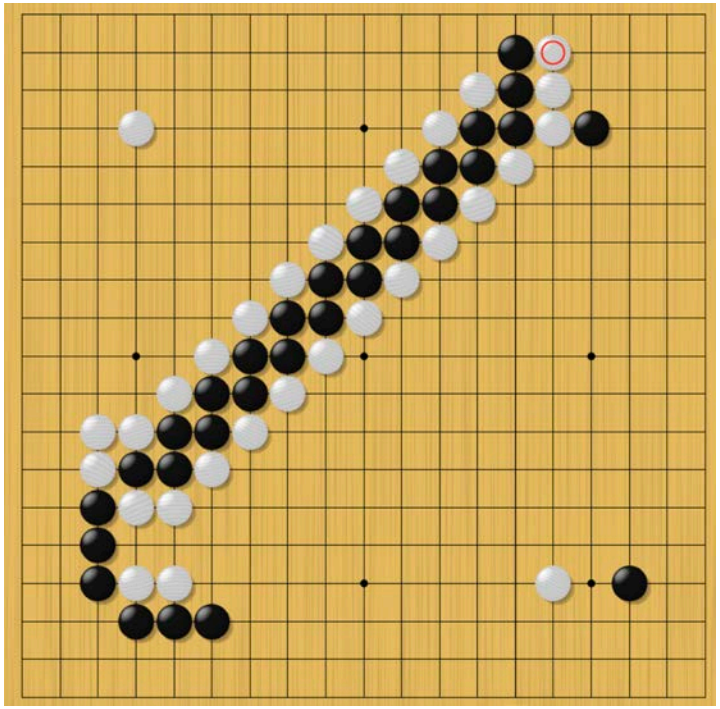
Run a ladder and lost



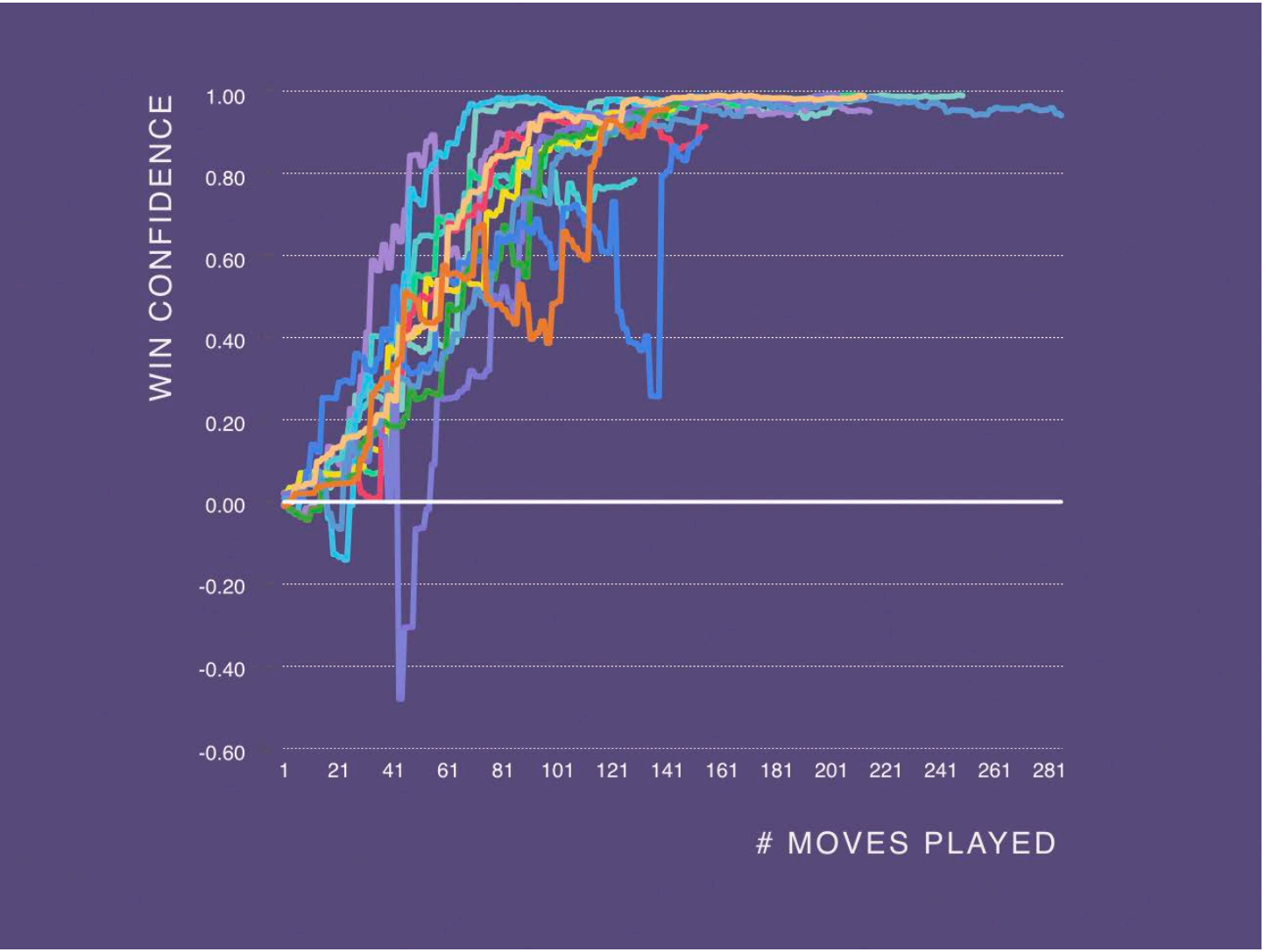
Run shorter ladder and lost



Doesn’t run ladder



Ladders are not “learned”, they are memorized.



20-0 against top Go players

How would a human beat the AI?

How would a human beat the AI?