Digital Image Analysis: Image Pattern Classification

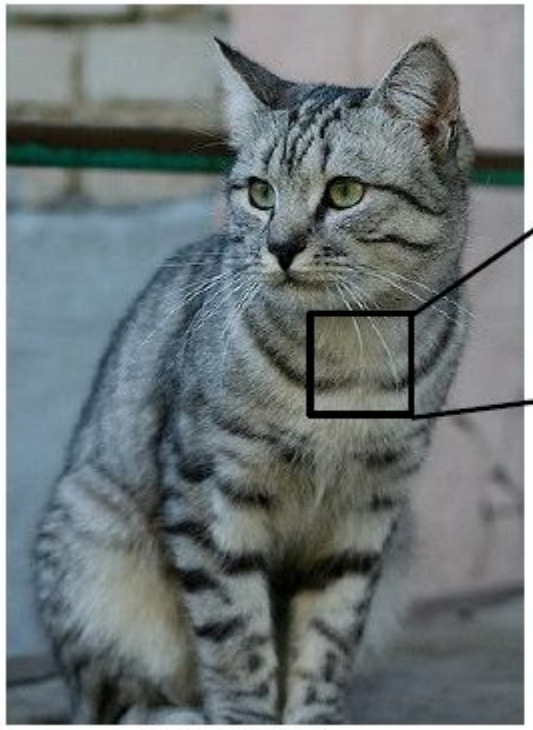
Image Classification: A core task in Computer Vision

(assume given set of discrete labels)
{dog, cat, truck, plane, ...}

cat

Slide Credit: Fei Fei Li et al., Stanford

The Problem: Semantic Gap




```
[[105 112 108 111 104  99 106  99  96 103 112 119 104  97  93  87]
 [ 91  98 102 106 104  79  98 103  99 105 123 136 118 105  94  85]
 [ 76  85  90 105 120 105  87  96  95  99 115 112 106 103  99  85]
 [ 99  81  81  93 120 131 127 100  95  98 102  99  96  93 101  94]
 [106  91  61  64  69  91  88  85 101 107 109  98  75  84  96  95]
 [114 108  85  55  55  69  64  54  64  87 112 129  98  74  84  91]
 [133 137 147 103  65  81  88  65  52  54  74  84 102  93  85  82]
 [120 137 144 140 109  95  86  70  62  65  63  63  68  73  86 101]
 [125 133 148 137 119 121 117  94  65  79  80  65  54  64  72  98]
 [127 125 131 147 133 127 126 131 111  96  89  75  61  64  72  84]
 [115 114 109 123 150 148 131 118 113 109 100  92  74  65  72  78]
 [ 89  93  90  97 108 147 131 118 113 114 113 109 106  95  77  80]
 [ 63  77  86  81  77  79 102 123 117 115 117 125 125 130 115  87]
 [ 62  65  82  89  78  71  88 101 124 126 119 101 107 114 131 119]
 [ 63  65  75  88  89  71  62  81 120 138 135 105  81  98 110 118]
 [ 87  65  71  87 106  95  69  45  76 130 126 107  92  94 105 112]
 [118  97  82  86 117 123 116  66  41  51  95  93  89  95 102 107]
 [164 146 112  80  82 120 124 104  76  48  45  66  88 101 102 109]
 [157 170 157 120  93  86 114 132 112  97  69  55  78  82  99  94]
 [130 128 134 161 139 100 109 118 121 134 114  87  65  53  69  86]
 [120 112  96 117 150 144 128 115 104 107 102  93  87  81  72  79]
 [123 107  96  86  83 112 153 140 122 109 104  75  88 107 112  99]
 [122 121 102  80  82  86  94 117 145 148 153 102  58  78  92 107]
 [122 164 148 103  71  56  78  83  93 103 119 139 102  61  69  84]]
```

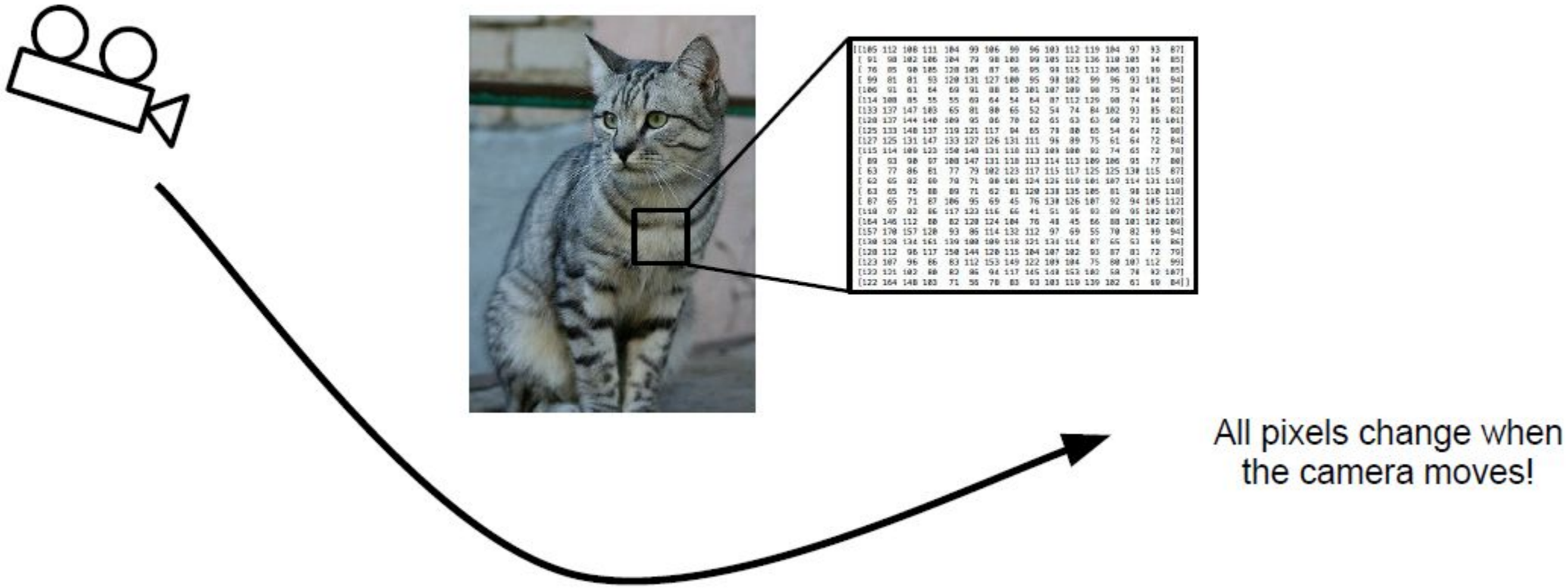
What the computer sees

An image is just a big grid of numbers between [0, 255]:

e.g. 800 x 600 x 3
(3 channels RGB)

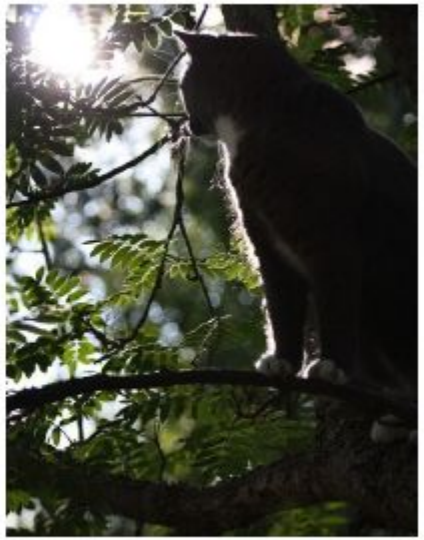
Slide Credit: Fei Fei Li et al., Stanford

Challenges: Viewpoint variation



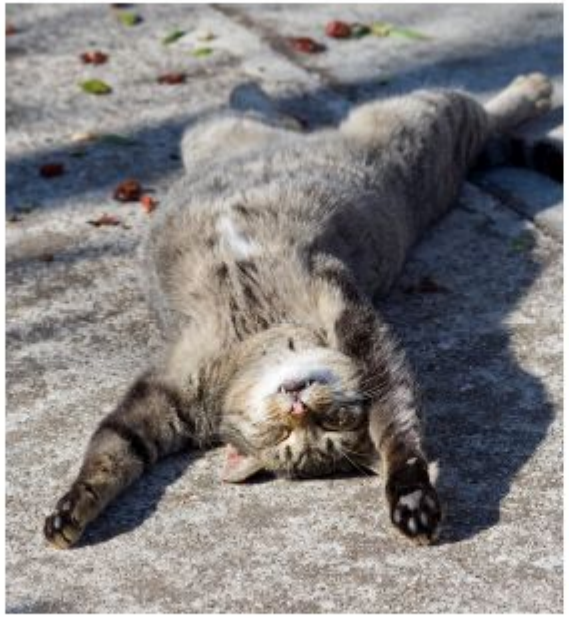
Slide Credit: Fei Fei Li et al., Stanford

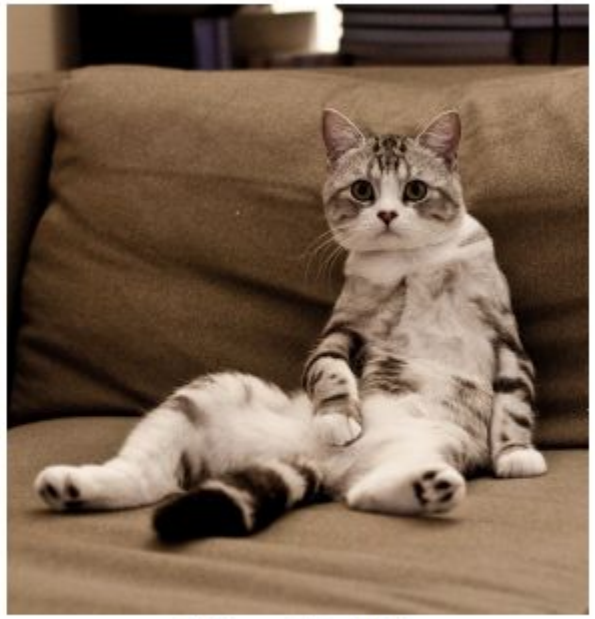
Challenges: Illumination



Slide Credit: Fei Fei Li et al., Stanford

Challenges: Deformation





Slide Credit: Fei Fei Li et al., Stanford

Challenges: Occlusion

This image is CC0 1.0 public domain

This image is CC0 1.0 public domain

This image by jonsson is licensed under CC-BY 2.0

Slide Credit: Fei Fei Li et al., Stanford

Challenges: Background Clutter

This image is CC0 1.0 public domain

This image is CC0 1.0 public domain

Slide Credit: Fei Fei Li et al., Stanford

Challenges: Intraclass variation

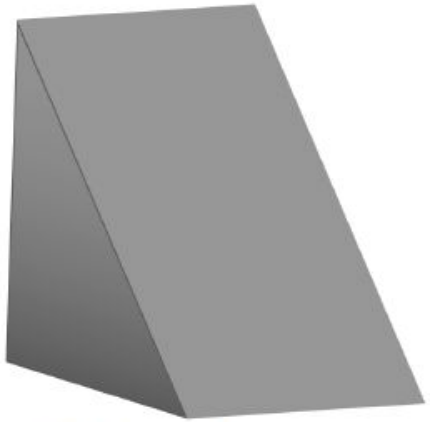
This image is CC0 1.0 public domain

Slide Credit: Fei Fei Li et al., Stanford

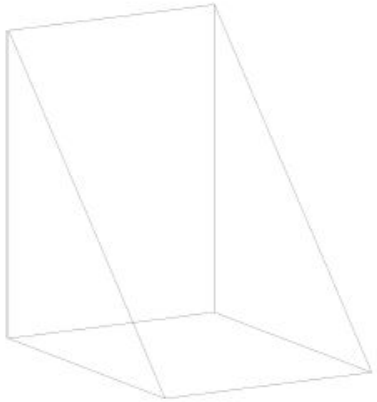
Brief History of Image Representation and Classification

Block world

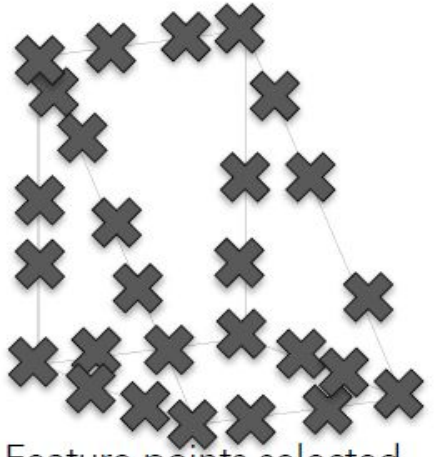
Larry Roberts, 1963



(a) Original picture



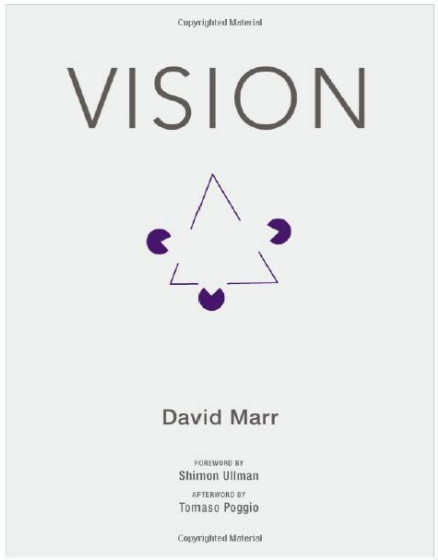
(b) Differentiated picture



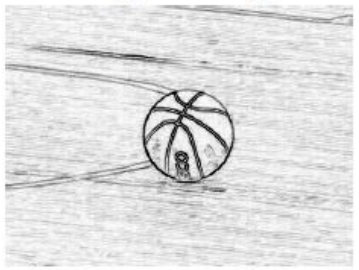
(c) Feature points selected

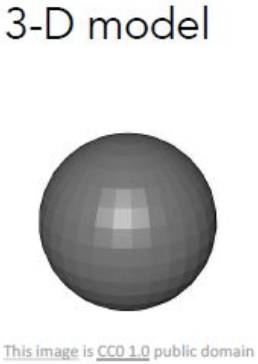
Slide Credit: Fei Fei Li et al., Stanford

Brief History of Image Representation and Classification

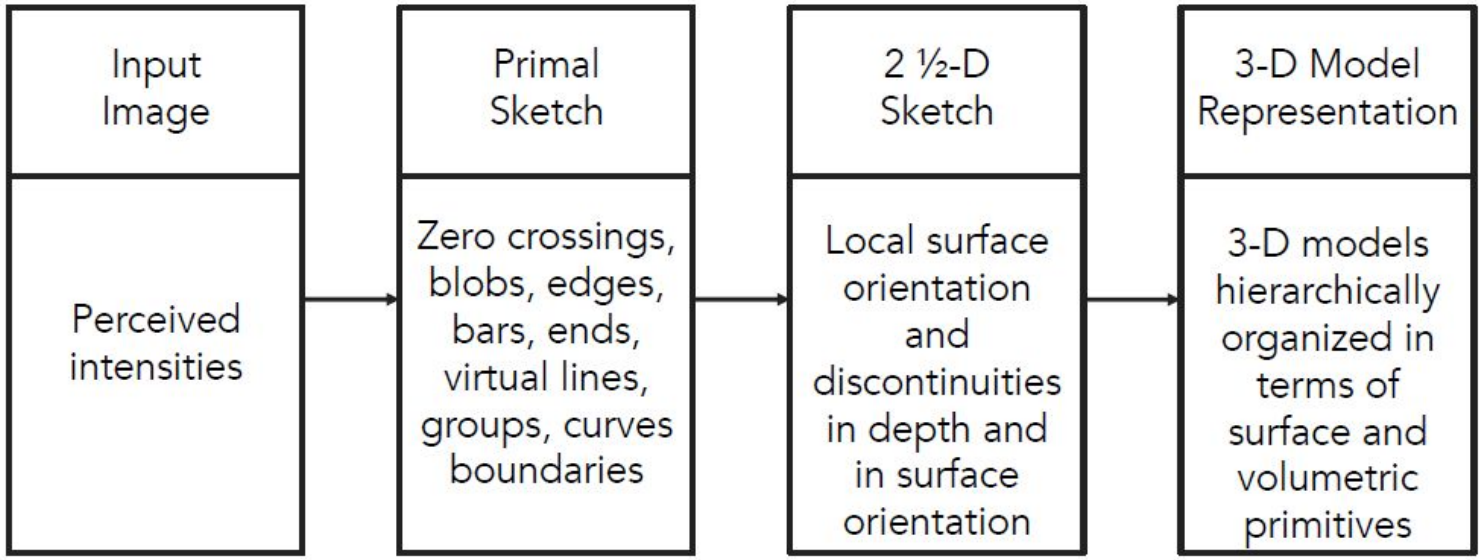


This image is CC0 1.0 public domain





This image is CC0 1.0 public domain



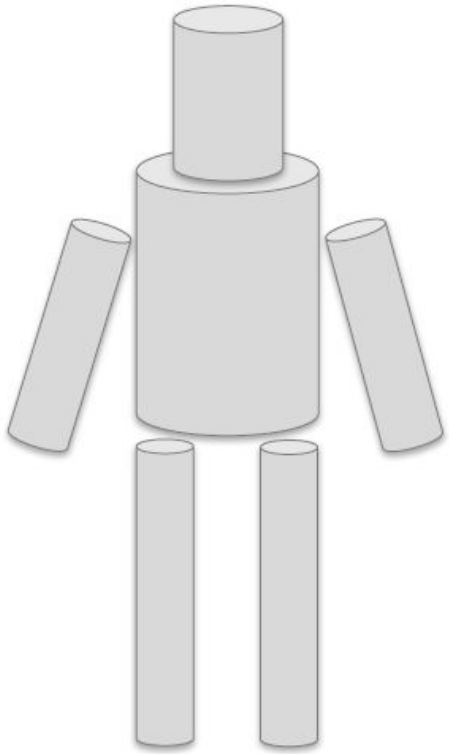
Stages of Visual Representation, David Marr, 1970s

Slide Credit: Fei Fei Li et al., Stanford

Brief History of Image Representation and Classification

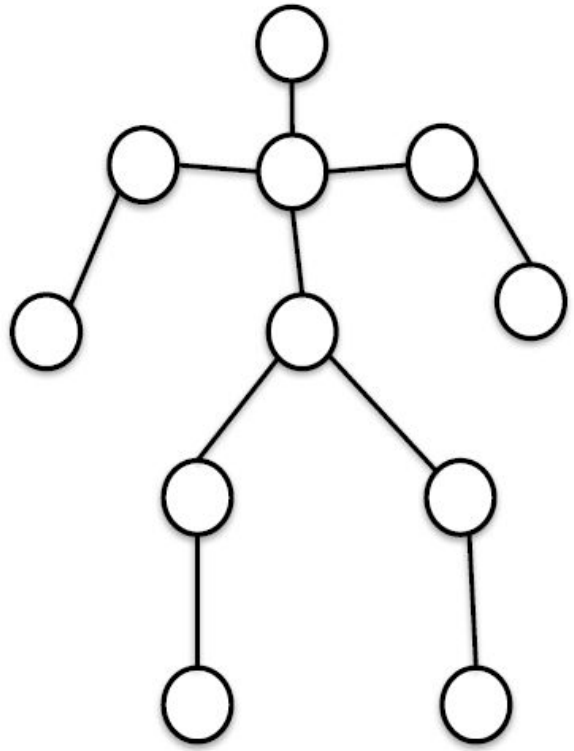
- Generalized Cylinder

Brooks & Binford, 1979



- Pictorial Structure

Fischler and Elschlager, 1973



Slide Credit: Fei Fei Li et al., Stanford

Brief History of Image Representation and Classification

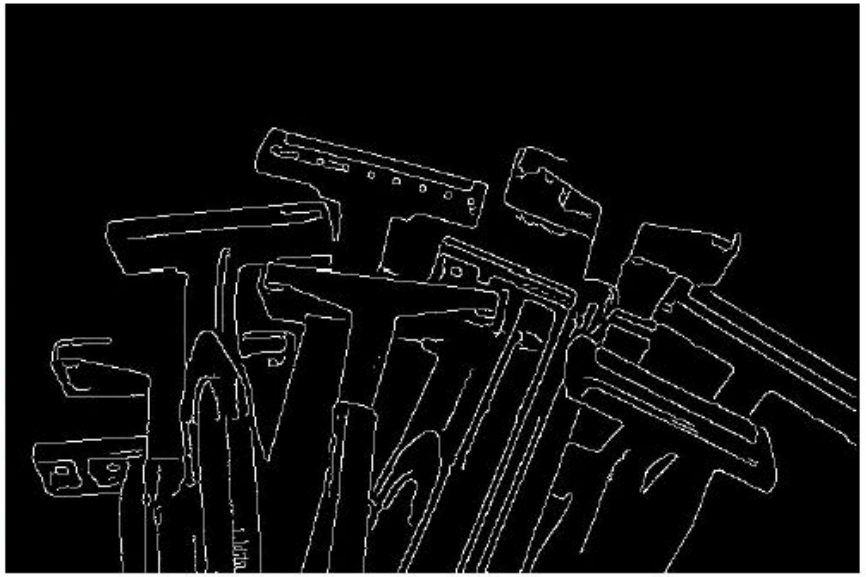


Image is CC0 1.0 public domain

David Lowe, 1987

Slide Credit: Fei Fei Li et al., Stanford

Brief History of Image Representation and Classification

Slide Credit: Fei Fei Li et al., Stanford

Brief History of Image Representation and Classification

Slide Credit: Fei Fei Li et al., Stanford

Brief History of Image Representation and Classification

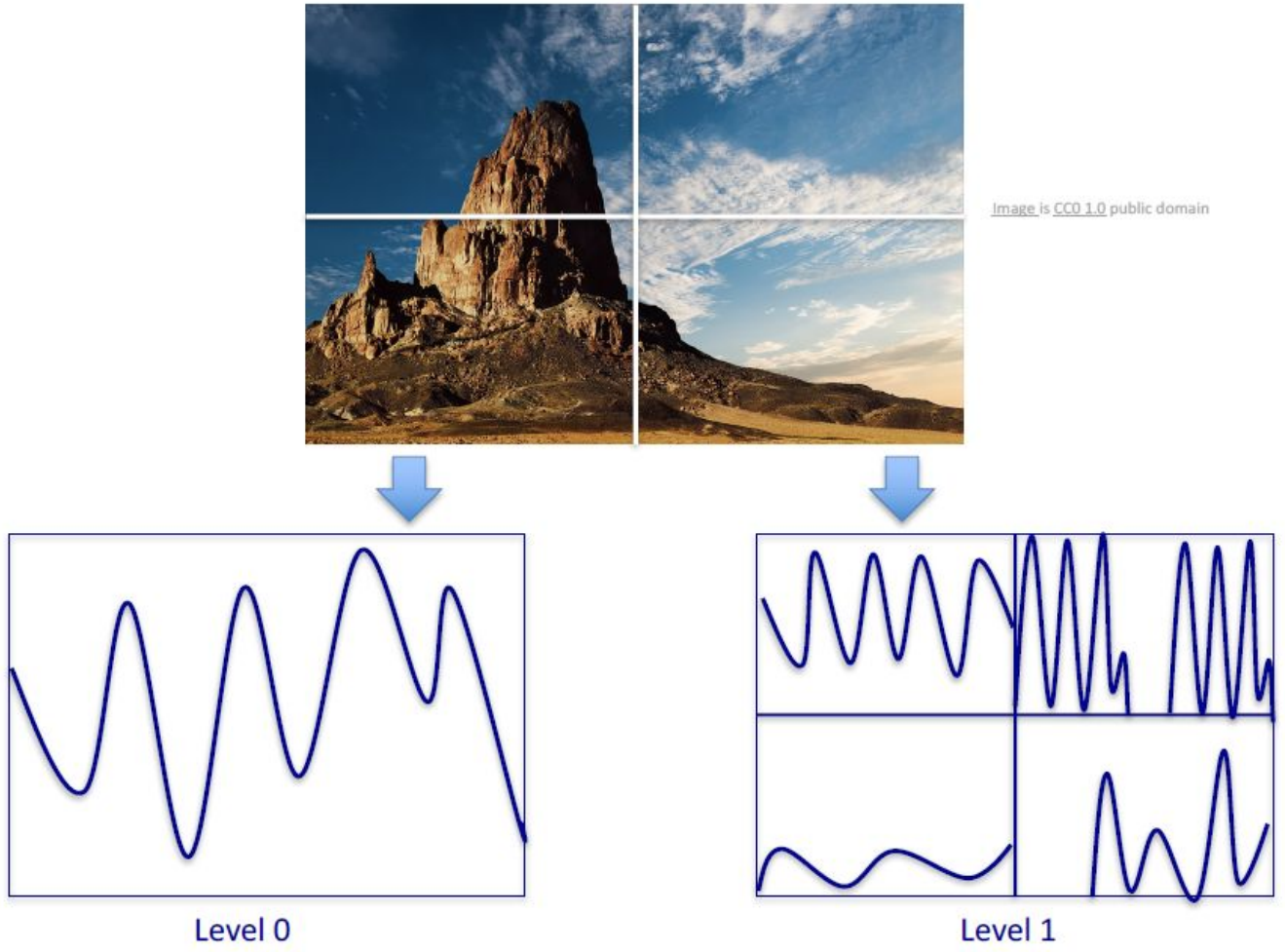
Image is public domain

Image is public domain

"SIFT" & Object Recognition, David Lowe, 1999

Slide Credit: Fei Fei Li et al., Stanford

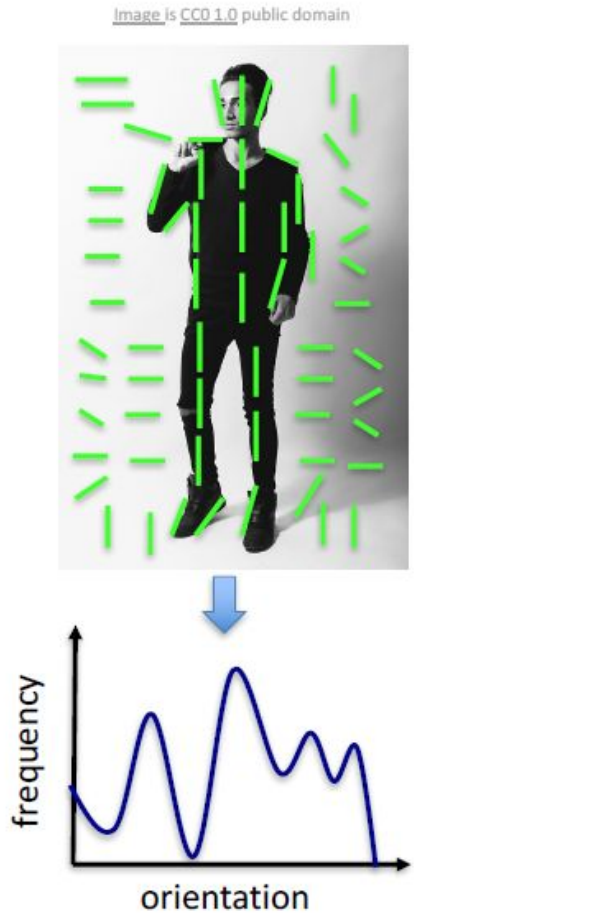
Brief History of Image Representation and Classification



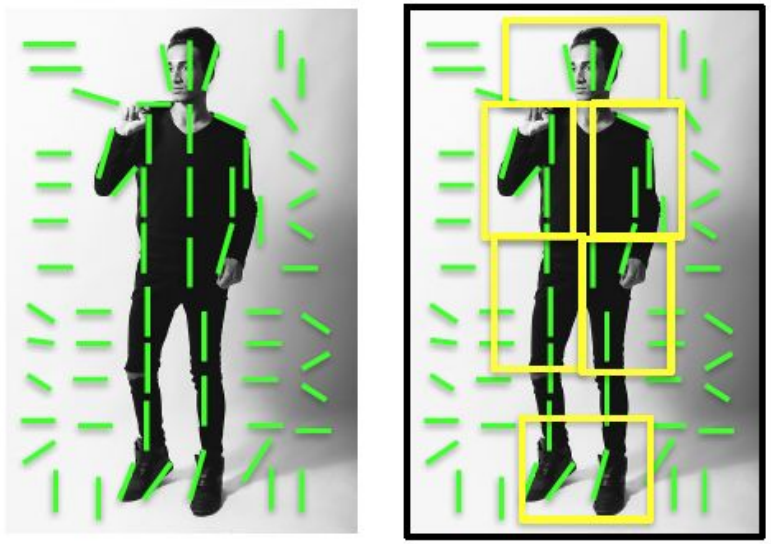
Spatial Pyramid Matching, Lazebnik, Schmid & Ponce, 2006

Slide Credit: Fei Fei Li et al., Stanford

Brief History of Image Representation and Classification



Histogram of Gradients (HoG)
Dalal & Triggs, 2005

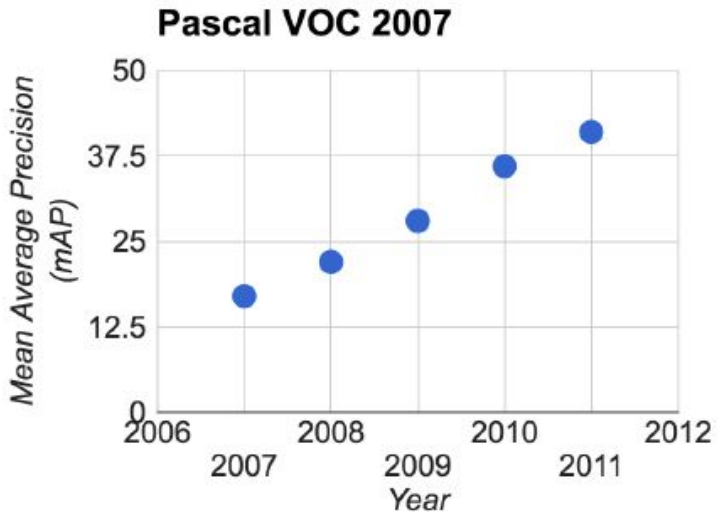


Deformable Part Model
Felzenswalb, McAllester, Ramanan, 2009

Slide Credit: Fei Fei Li et al., Stanford

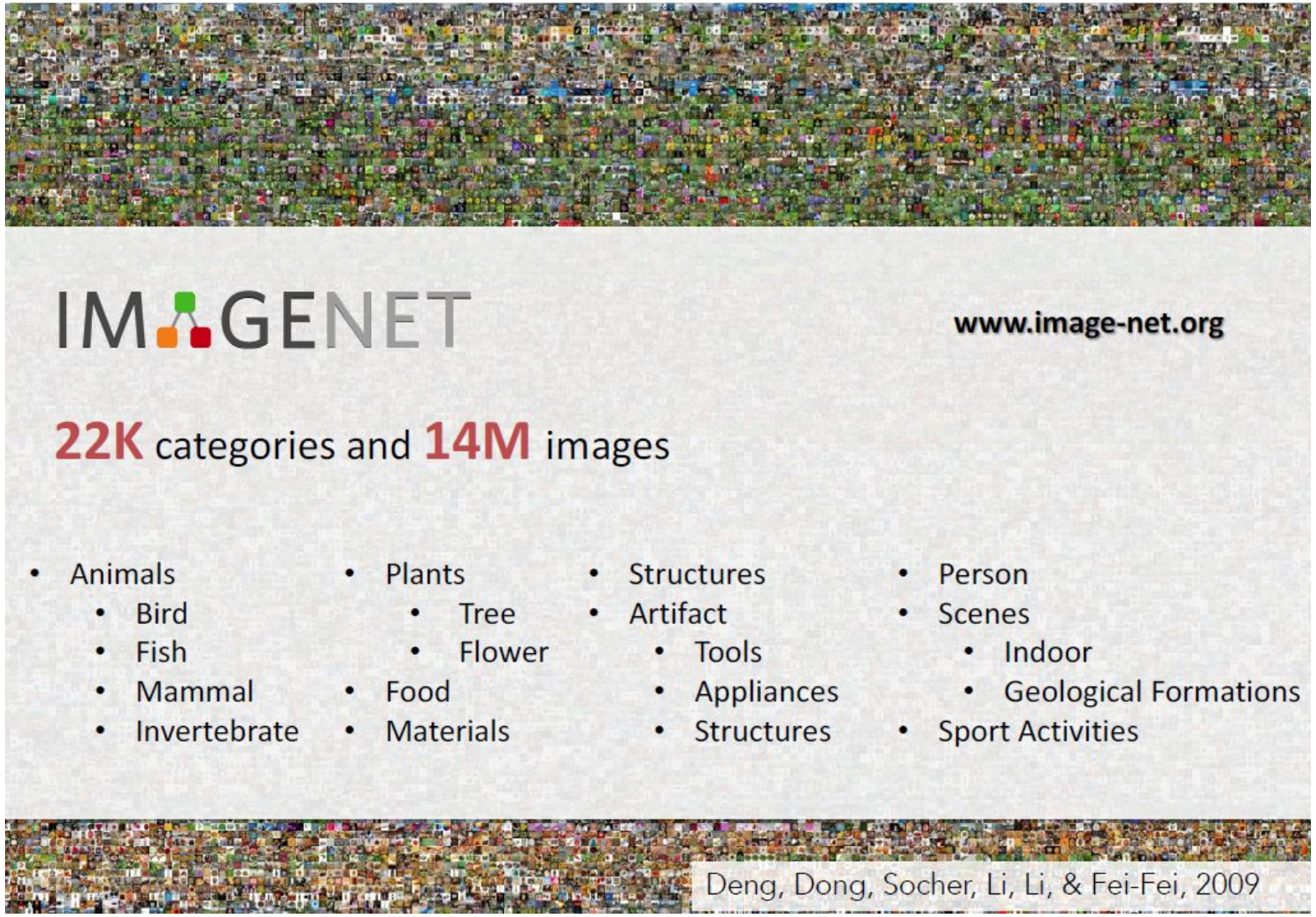
Brief History of Image Representation and Classification

PASCAL Visual Object Challenge
(20 object categories)
[Everingham et al. 2006-2012]



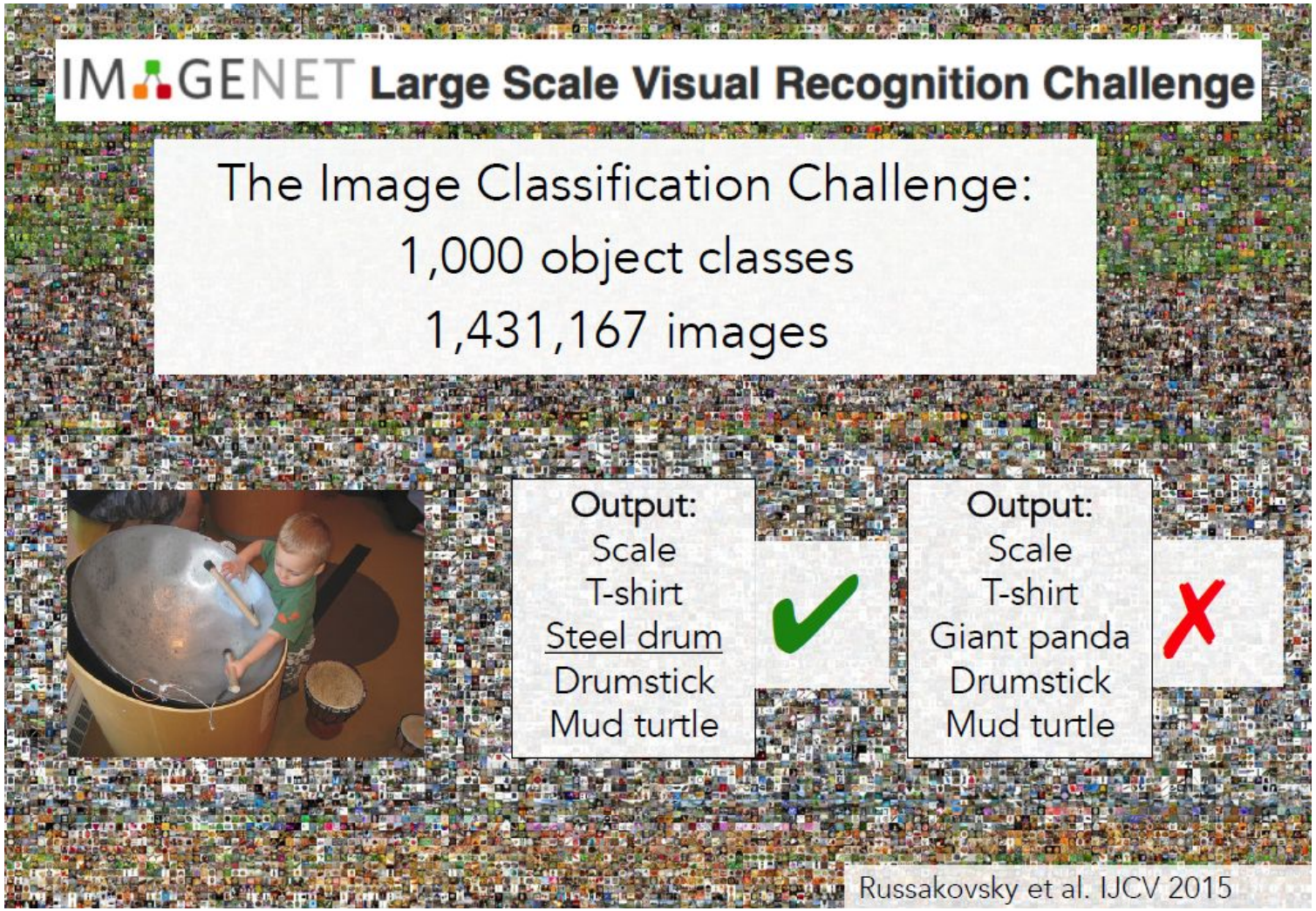
Slide Credit: Fei Fei Li et al., Stanford

Brief History of Image Representation and Classification



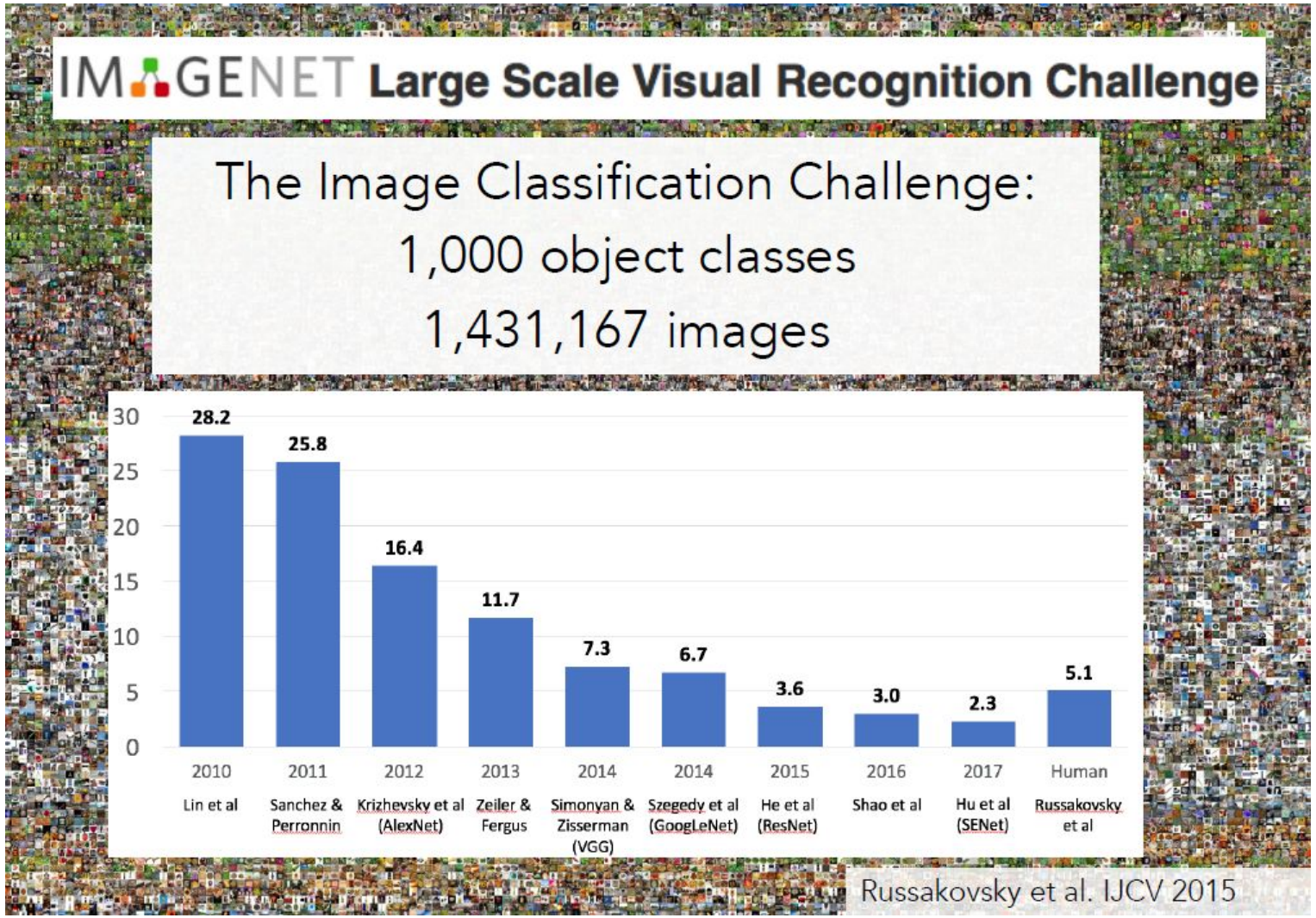
Slide Credit: Fei Fei Li et al., Stanford

Brief History of Image Representation and Classification



Slide Credit: Fei Fei Li et al., Stanford

Brief History of Image Representation and Classification



Slide Credit: Fei Fei Li et al., Stanford

Convolutional Neural Networks (CNN) have become an important tool for object recognition

Slide Credit: Fei Fei Li et al., Stanford

IMAGENET Large Scale Visual Recognition Challenge

Year 2010

NEC-UIUC

Dense descriptor grid: HOG, LBP

Coding: local coordinate, super-vector

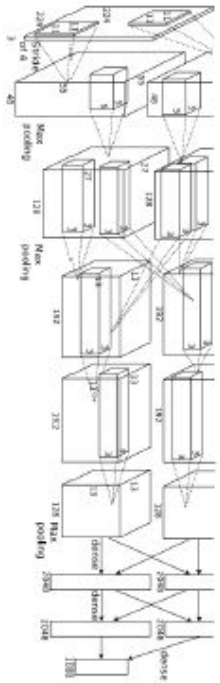
Pooling, SPM

Linear SVM

[Lin CVPR 2011]

Year 2012

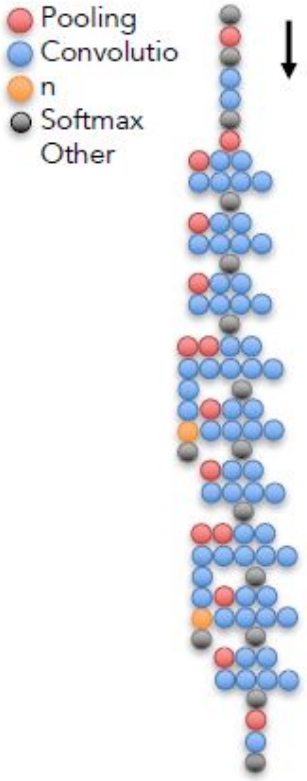
SuperVision



[Krizhevsky NIPS 2012]

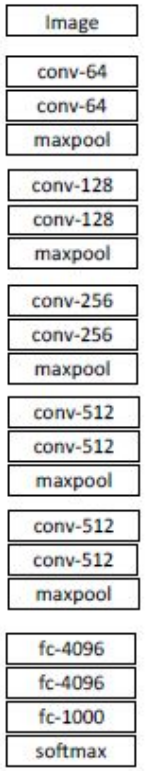
Year 2014

GoogLeNet



[Szegedy arxiv 2014]

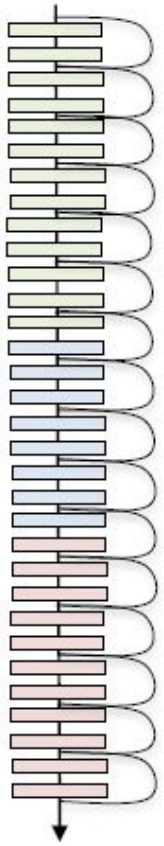
VGG



[Simonyan arxiv 2014]

Year 2015

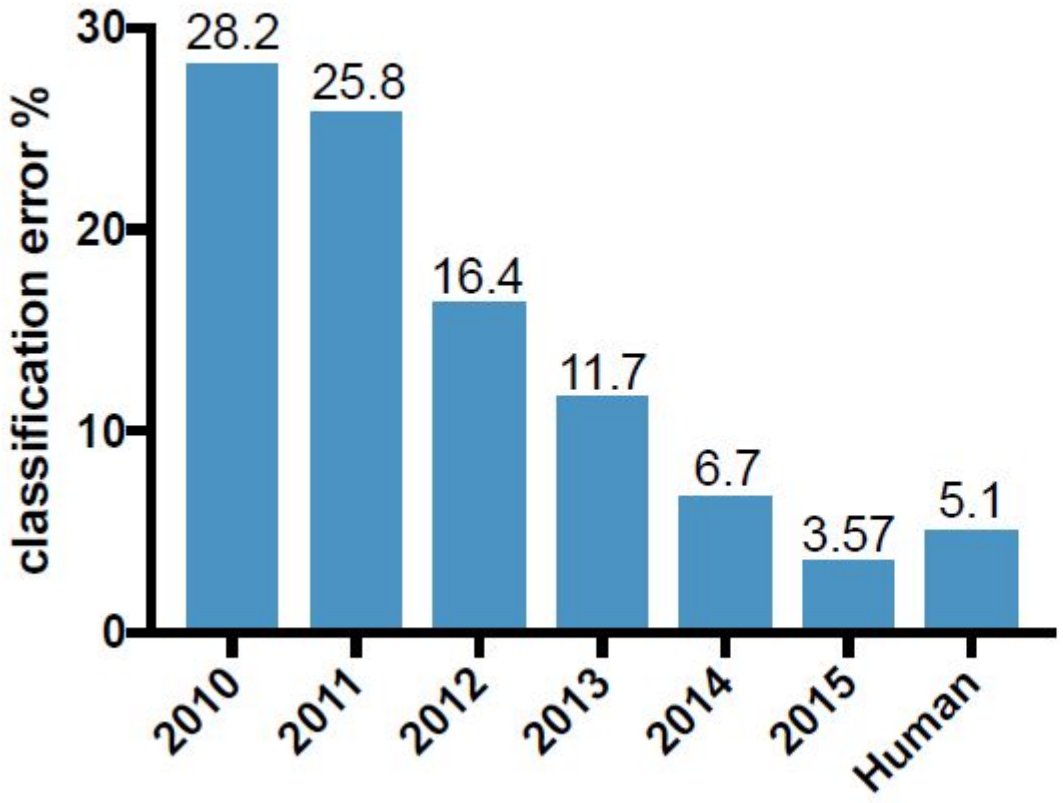
MSRA



[He ICCV 2015]

Slide Credit: Fei-Fei Li et al., Stanford

ImageNet Challenge: Classification Task



2012: AlexNet. First CNN to win.
- 8 layers, 61 million parameters

2013: ZFNet
- 8 layers, more filters

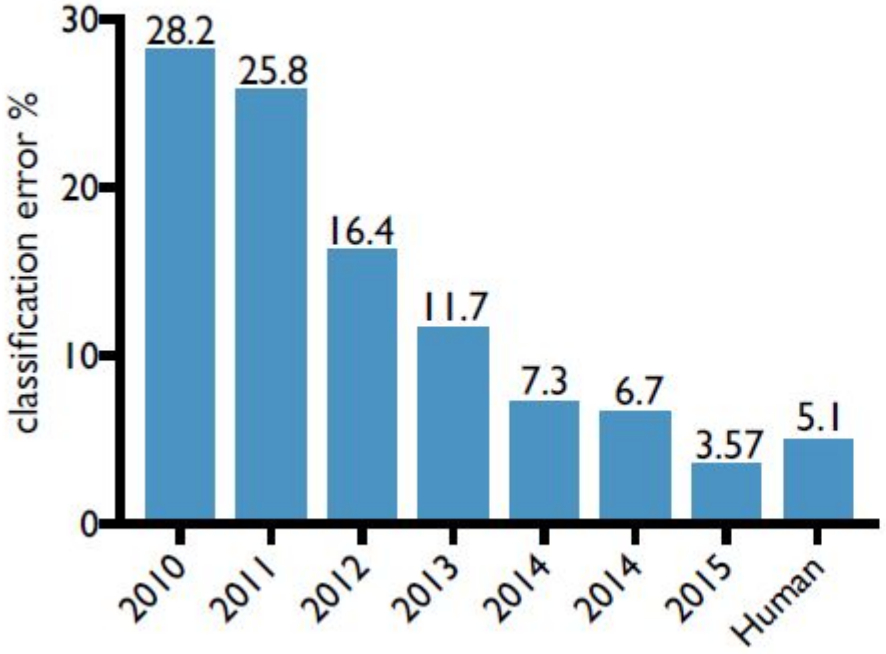
2014: VGG
- 19 layers

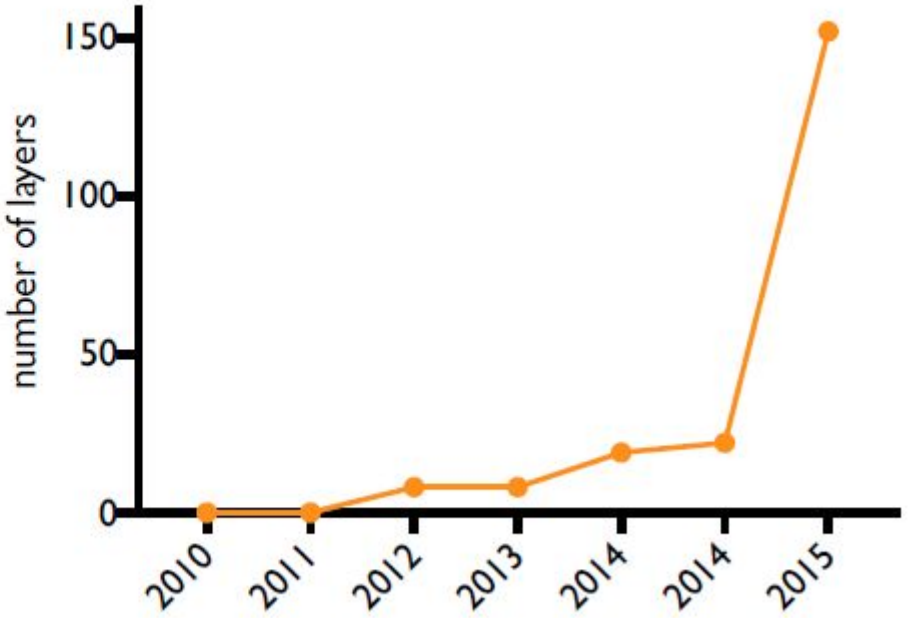
2014: GoogLeNet
- "Inception" modules
- 22 layers, 5million parameters

2015: ResNet
- 152 layers

Slide Credit: Ava Soleimany, MIT

ImageNet Challenge: Classification Task



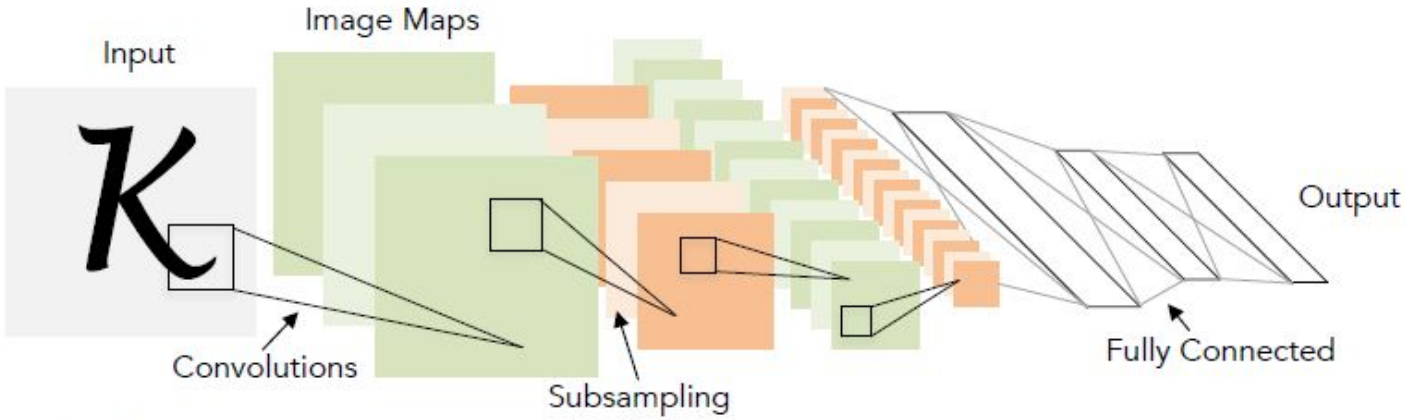


Slide Credit: Ava Soleimany, MIT

Convolutional Neural Networks (CNN)
were not invented overnight

Slide Credit: Fei Fei Li et al., Stanford

1998
LeCun et al.



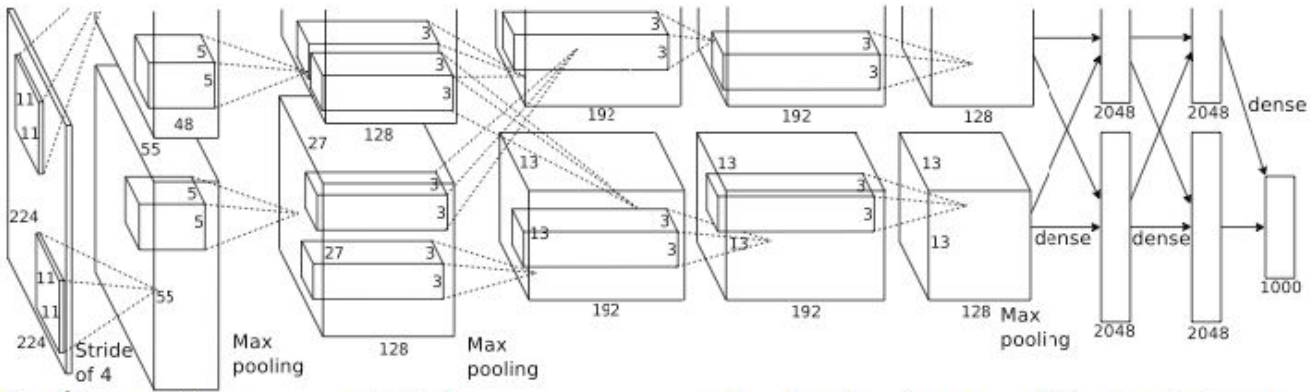
# of transistors

10^6

# of pixels used in training

10^7 NIST

2012
Krizhevsky et al.



# of transistors

10^9

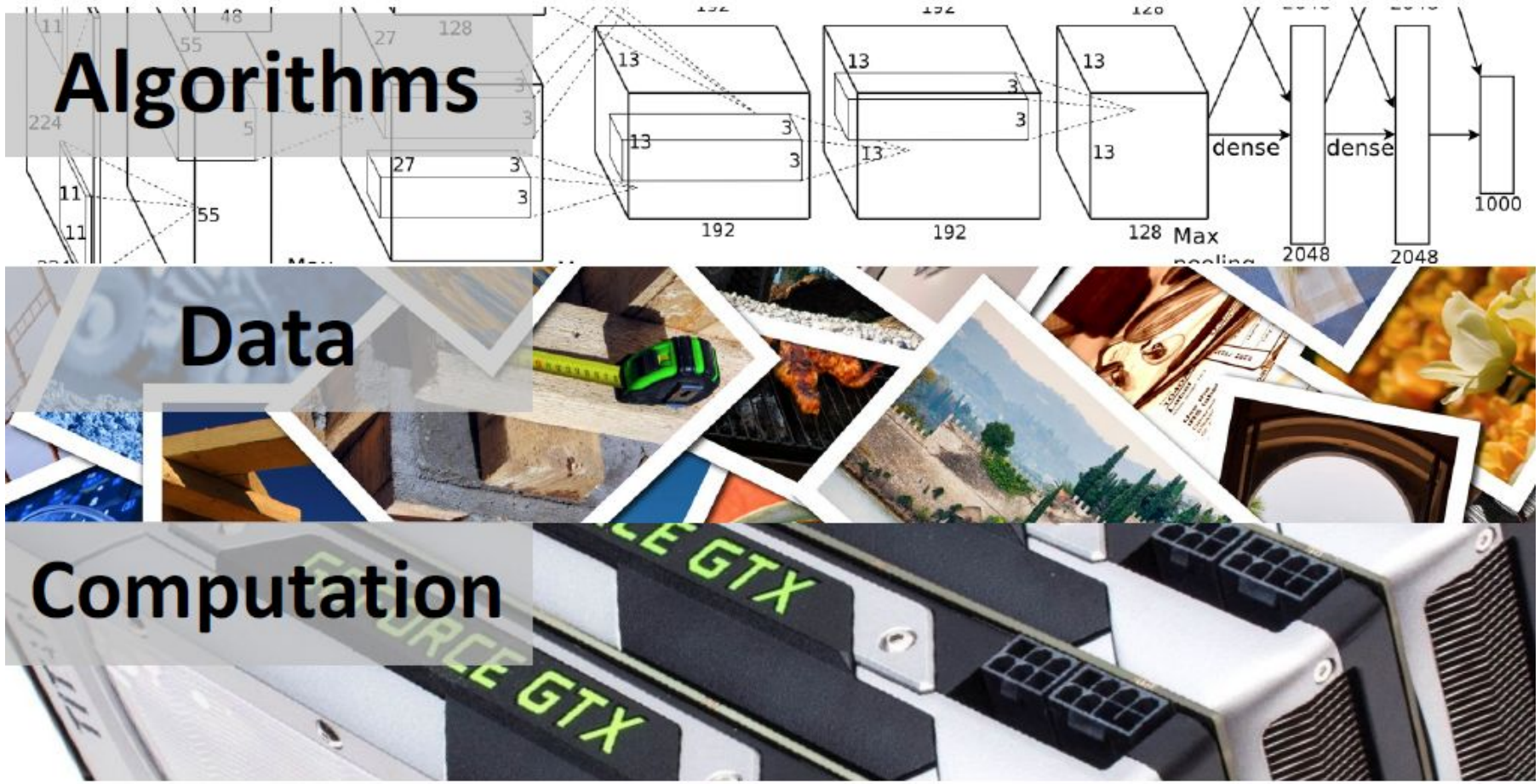
GPUs

# of pixels used in training

10^{14} IMAGENET

Slide Credit: Fei Fei Li et al., Stanford

Ingredients for Deep Learning



Slide Credit: Fei Fei Li et al., Stanford

CNN Architectures for Image Classification

CNN Architectures

Case Studies

- AlexNet
- VGG
- GoogLeNet
- ResNet

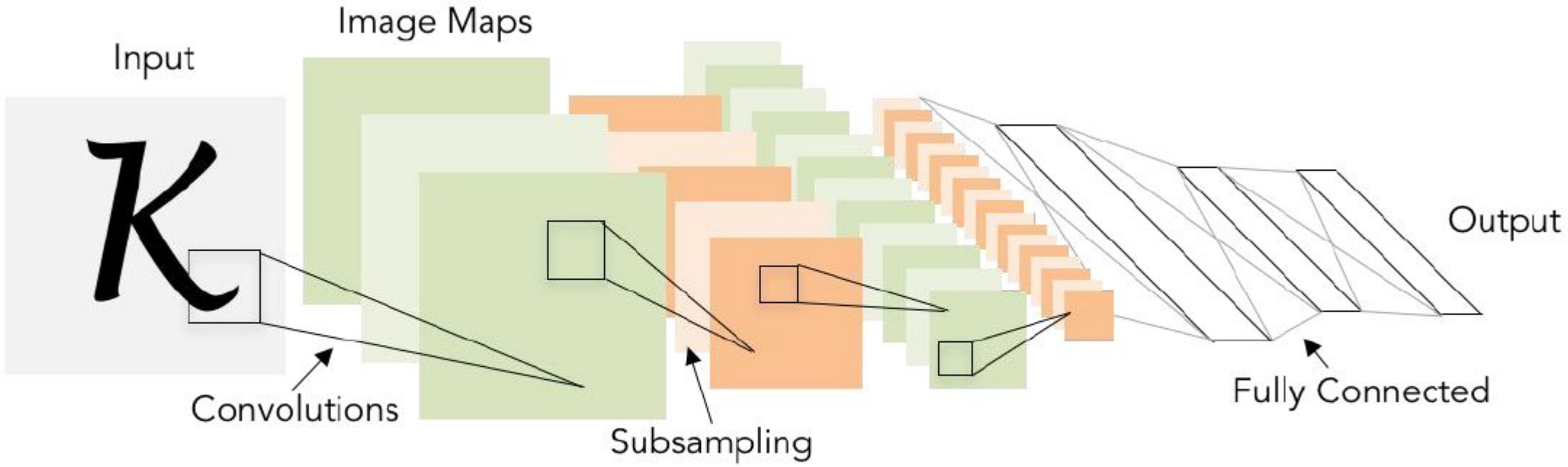
Also....

- NiN (Network in Network)
- Wide ResNet
- ResNeXT
- Stochastic Depth
- Squeeze-and-Excitation Network
- DenseNet
- FractalNet
- SqueezeNet
- NASNet

Slide Credit: Fei Fei Li et al., Stanford

Review: LeNet-5

[LeCun et al., 1998]



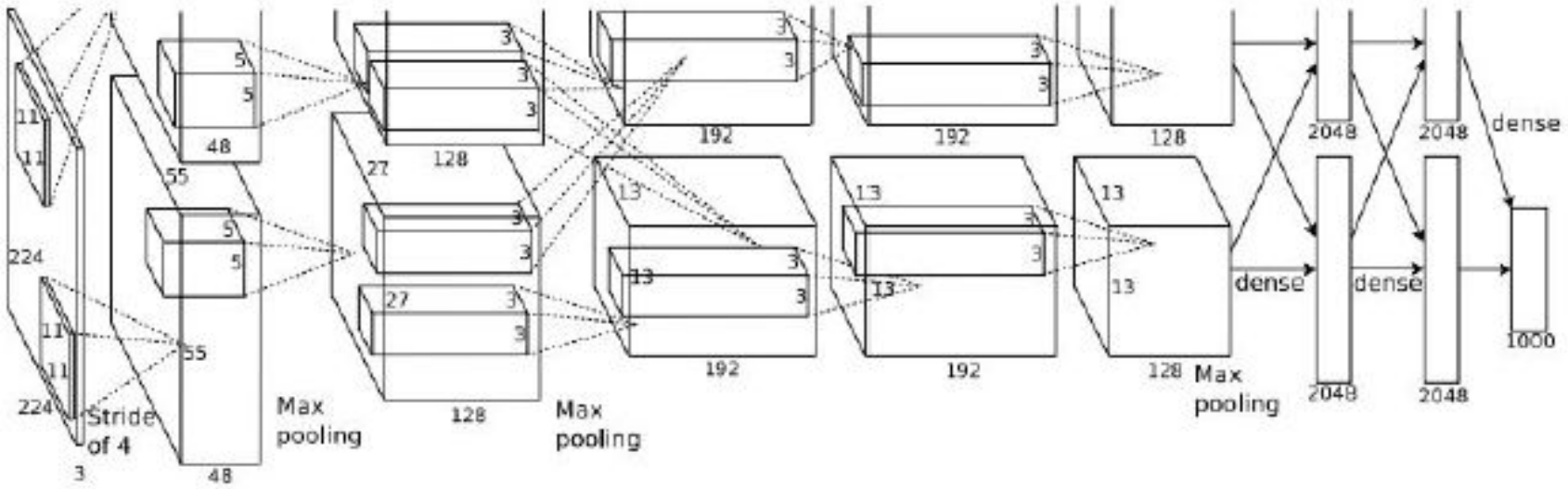
Conv filters were 5x5, applied at stride 1
Subsampling (Pooling) layers were 2x2 applied at stride 2
i.e. architecture is [CONV-POOL-CONV-POOL-FC-FC]

Slide Credit: Fei Fei Li et al., Stanford

Case Study: AlexNet

[Krizhevsky et al. 2012]

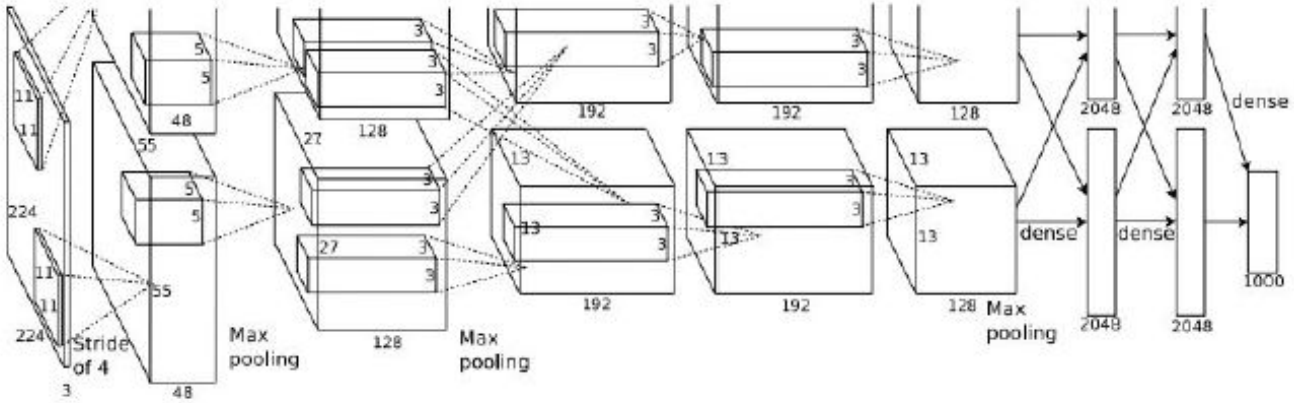
Architecture:
CONV1
MAX POOL1
NORM1
CONV2
MAX POOL2
NORM2
CONV3
CONV4
CONV5
Max POOL3
FC6
FC7
FC8



Slide Credit: Fei Fei Li et al., Stanford

Case Study: AlexNet

[Krizhevsky et al. 2012]

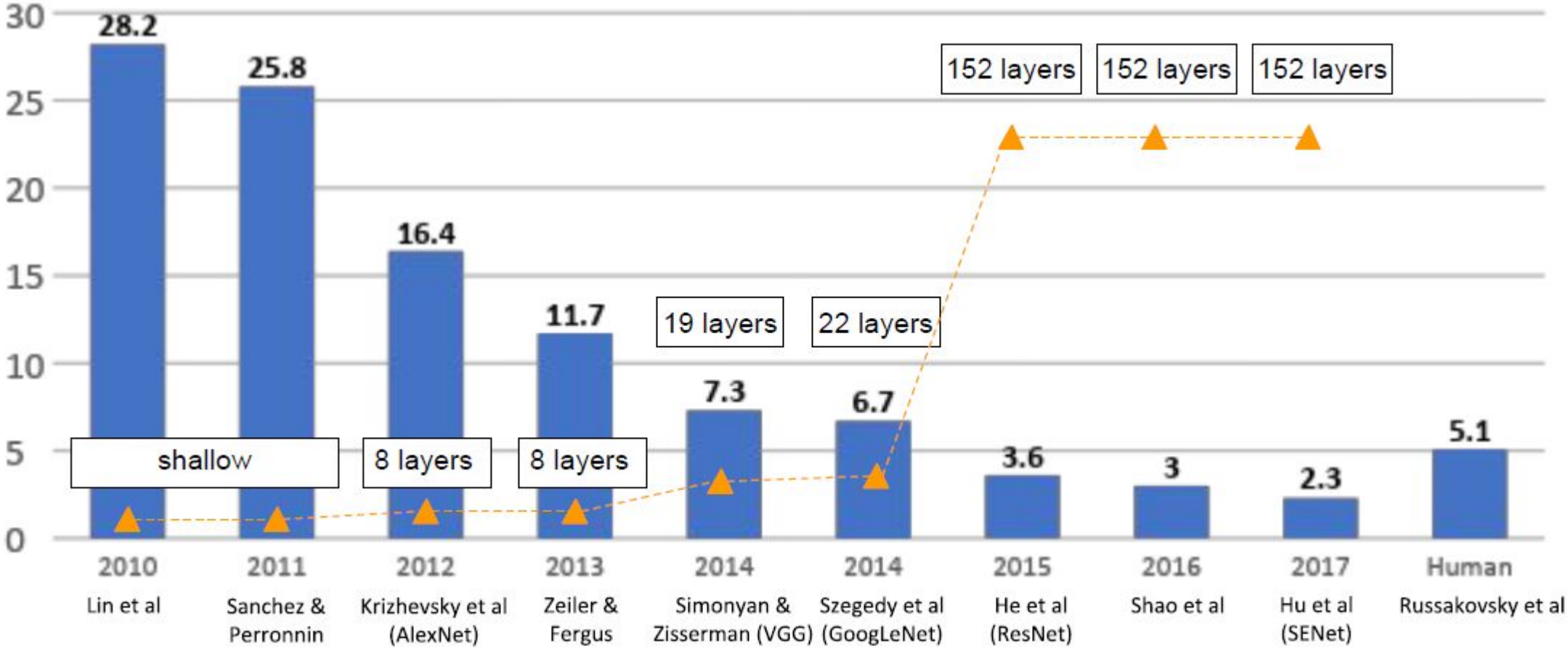


Full (simplified) AlexNet architecture:
[227x227x3] INPUT
[55x55x96] CONV1: 96 11x11 filters at stride 4, pad 0
[27x27x96] MAX POOL1: 3x3 filters at stride 2
[27x27x96] NORM1: Normalization layer
[27x27x256] CONV2: 256 5x5 filters at stride 1, pad 2
[13x13x256] MAX POOL2: 3x3 filters at stride 2
[13x13x256] NORM2: Normalization layer
[13x13x384] CONV3: 384 3x3 filters at stride 1, pad 1
[13x13x384] CONV4: 384 3x3 filters at stride 1, pad 1
[13x13x256] CONV5: 256 3x3 filters at stride 1, pad 1
[6x6x256] MAX POOL3: 3x3 filters at stride 2
[4096] FC6: 4096 neurons
[4096] FC7: 4096 neurons
[1000] FC8: 1000 neurons (class scores)

Details/Retrospectives:
- first use of ReLU
- used Norm layers (not common anymore)
- heavy data augmentation
- dropout 0.5
- batch size 128
- SGD Momentum 0.9
- Learning rate 1e-2, reduced by 10 manually when val accuracy plateaus
- L2 weight decay 5e-4
- 7 CNN ensemble: 18.2% -> 15.4%

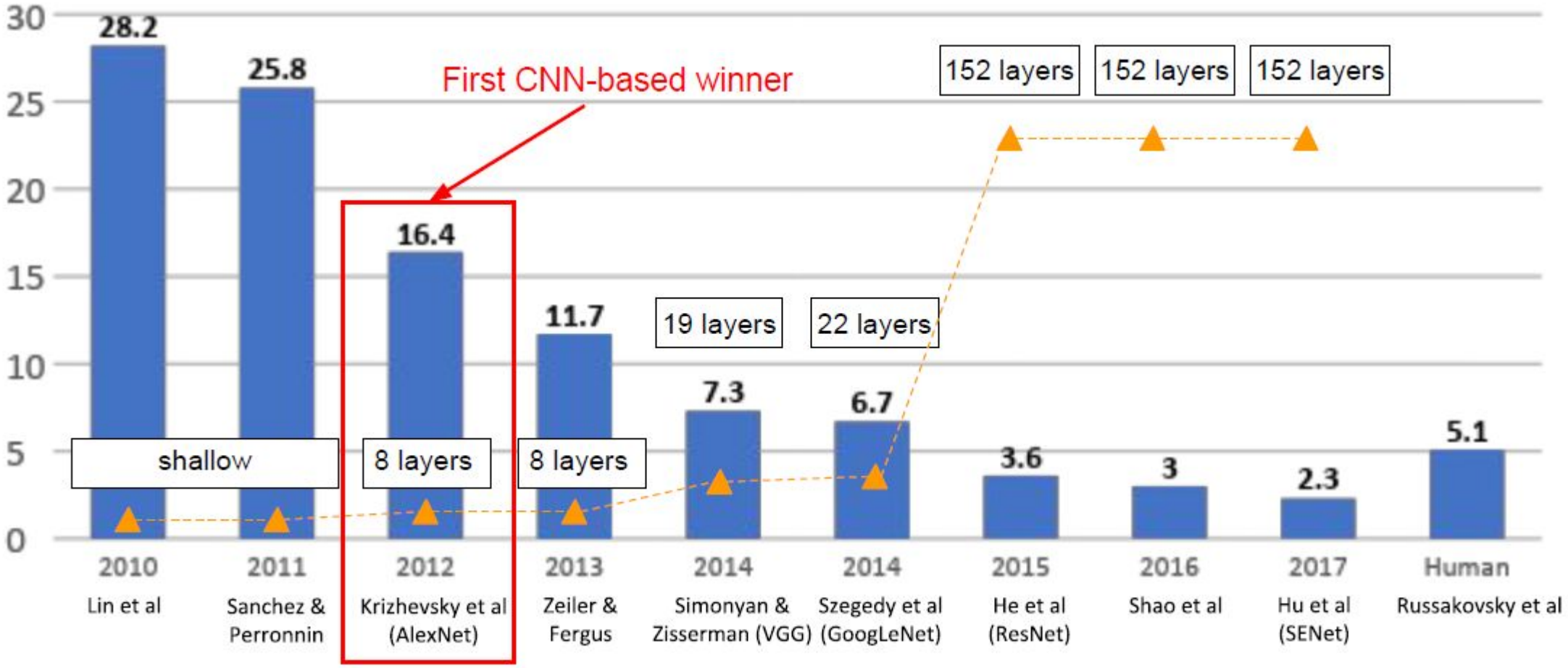
Slide Credit: Fei Fei Li et al., Stanford

ImageNet Large Scale Visual Recognition Challenge (ILSVRC) winners



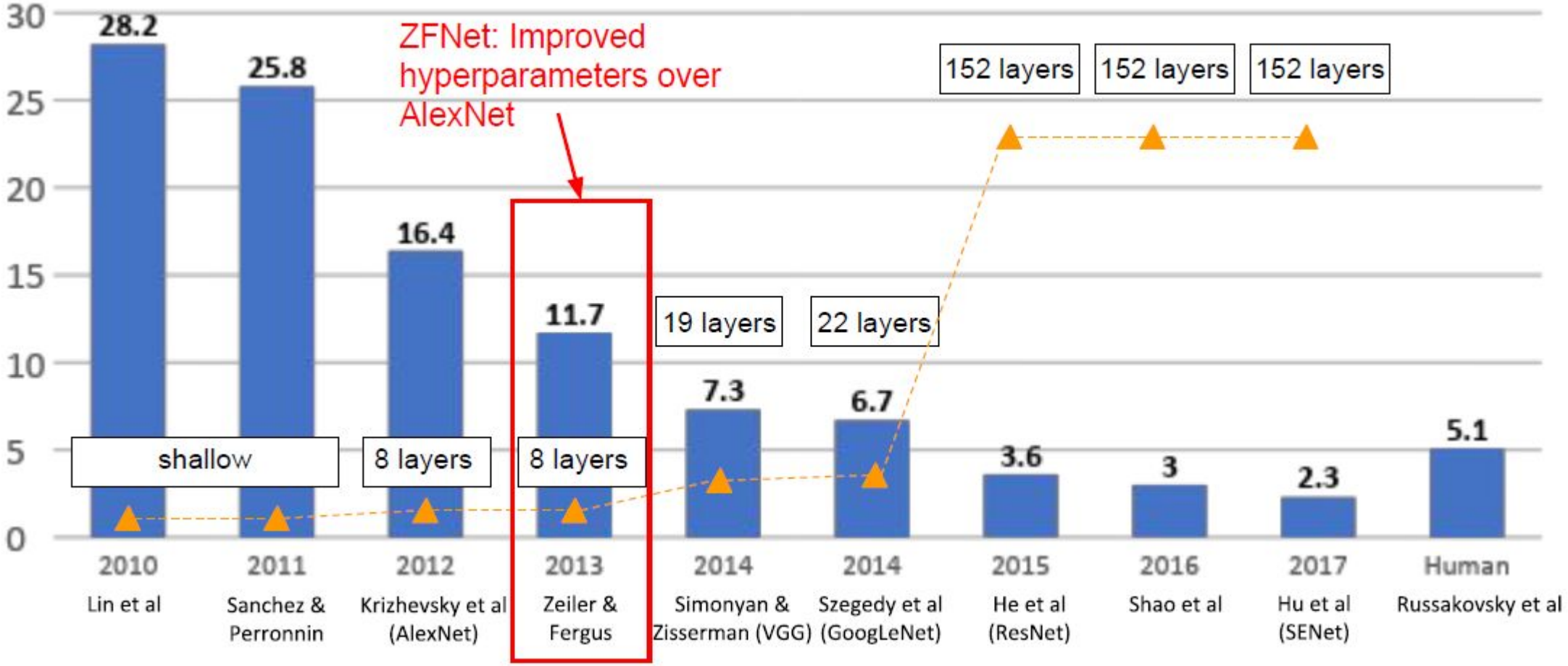
Slide Credit: Fei Fei Li et al., Stanford

ImageNet Large Scale Visual Recognition Challenge (ILSVRC) winners



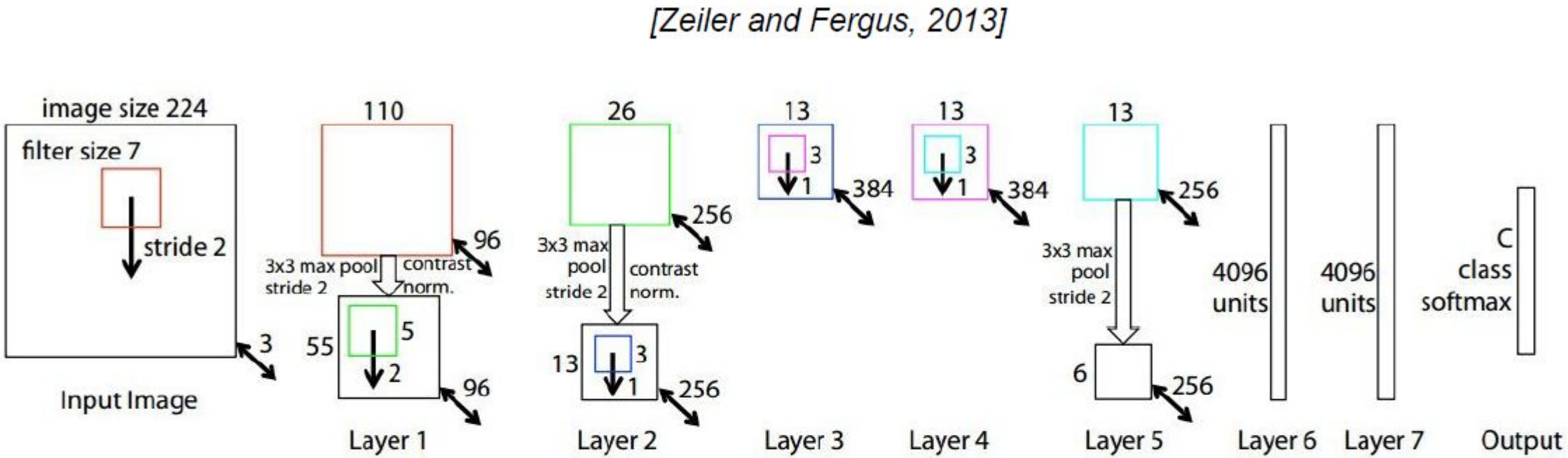
Slide Credit: Fei Fei Li et al., Stanford

ImageNet Large Scale Visual Recognition Challenge (ILSVRC) winners



Slide Credit: Fei Fei Li et al., Stanford

ZFNet

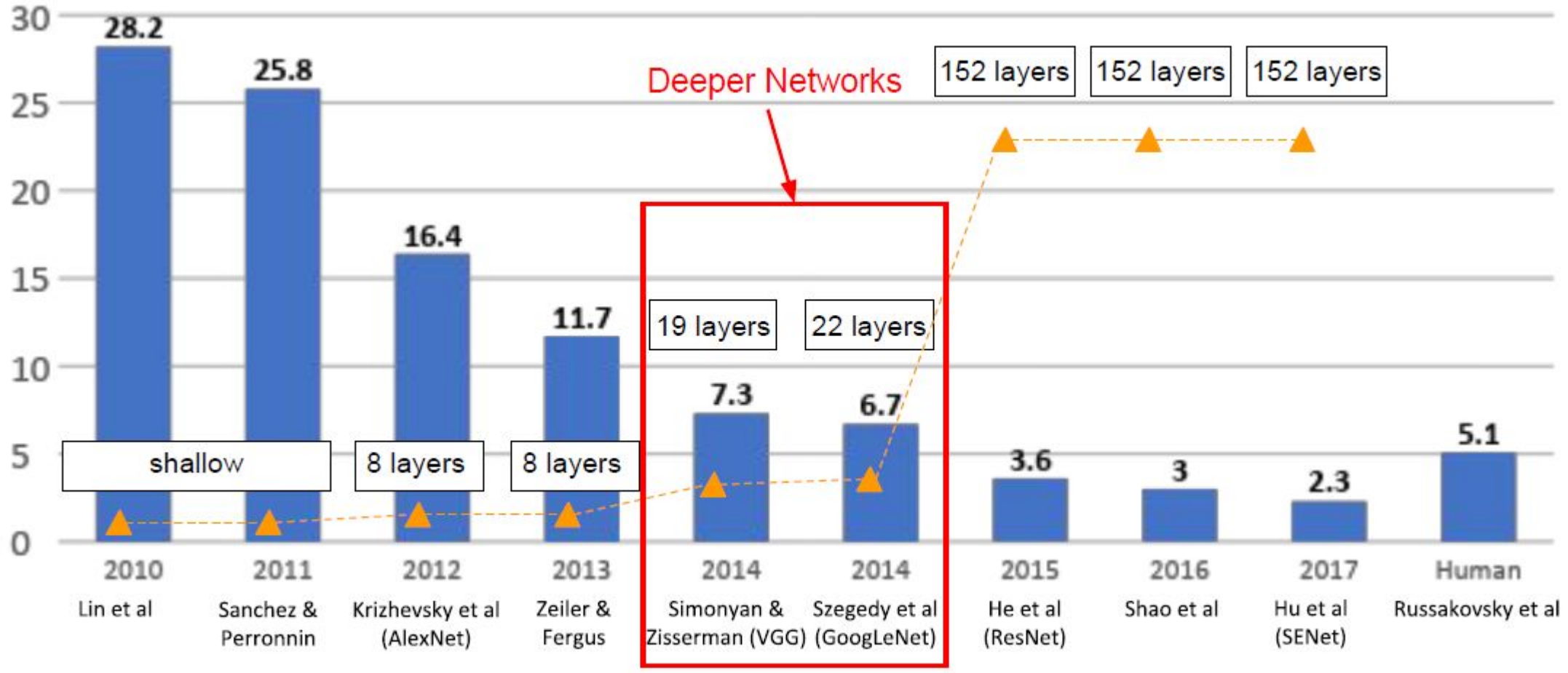


AlexNet but:
CONV1: change from (11x11 stride 4) to (7x7 stride 2)
CONV3,4,5: instead of 384, 384, 256 filters use 512, 1024, 512

ImageNet top 5 error: 16.4% -> 11.7%

Slide Credit: Fei Fei Li et al., Stanford

ImageNet Large Scale Visual Recognition Challenge (ILSVRC) winners



Slide Credit: Fei Fei Li et al., Stanford

Case Study: VGGNet

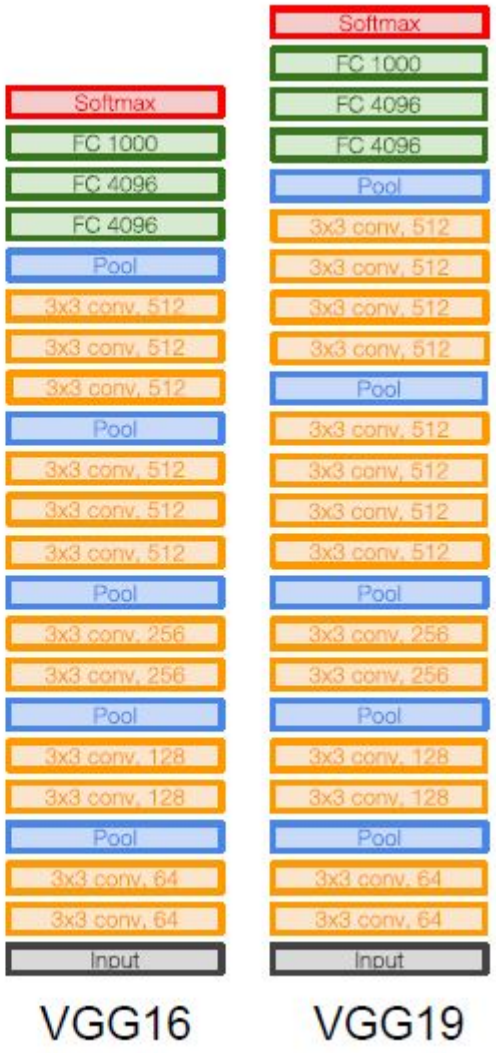
[Simonyan and Zisserman, 2014]

Small filters, Deeper networks

8 layers (AlexNet)
-> 16 - 19 layers (VGG16Net)

Only 3x3 CONV stride 1, pad 1
and 2x2 MAX POOL stride 2

11.7% top 5 error in ILSVRC'13
(ZFNet)
-> 7.3% top 5 error in ILSVRC'14



Details:

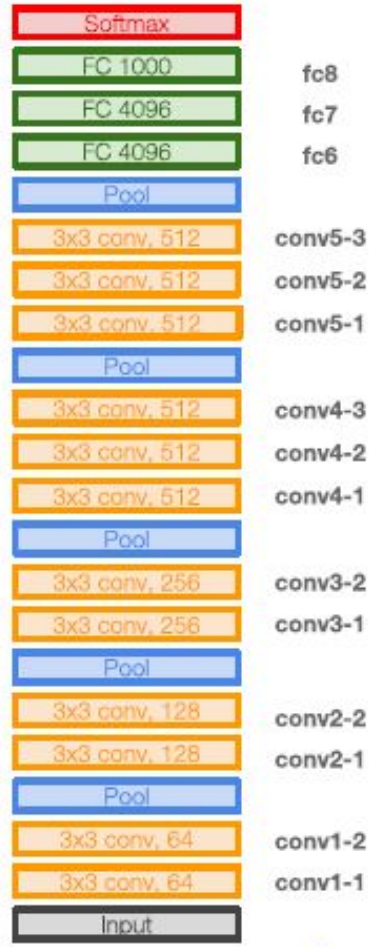
- ILSVRC'14 2nd in classification, 1st in localization
- Similar training procedure as Krizhevsky 2012
- No Local Response Normalisation (LRN)
- Use VGG16 or VGG19 (VGG19 only slightly better, more memory)
- Use ensembles for best results
- FC7 features generalize well to other tasks

Slide Credit: Fei Fei Li et al., Stanford

(not counting biases)

INPUT: [224x224x3] memory: 224*224*3=150K params: 0
CONV3-64: [224x224x64] memory: 224*224*64=3.2M params: (3*3*3)*64 = 1,728
CONV3-64: [224x224x64] memory: 224*224*64=3.2M params: (3*3*64)*64 = 36,864
POOL2: [112x112x64] memory: 112*112*64=800K params: 0
CONV3-128: [112x112x128] memory: 112*112*128=1.6M params: (3*3*64)*128 = 73,728
CONV3-128: [112x112x128] memory: 112*112*128=1.6M params: (3*3*128)*128 = 147,456
POOL2: [56x56x128] memory: 56*56*128=400K params: 0
CONV3-256: [56x56x256] memory: 56*56*256=800K params: (3*3*128)*256 = 294,912
CONV3-256: [56x56x256] memory: 56*56*256=800K params: (3*3*256)*256 = 589,824
CONV3-256: [56x56x256] memory: 56*56*256=800K params: (3*3*256)*256 = 589,824
POOL2: [28x28x256] memory: 28*28*256=200K params: 0
CONV3-512: [28x28x512] memory: 28*28*512=400K params: (3*3*256)*512 = 1,179,648
CONV3-512: [28x28x512] memory: 28*28*512=400K params: (3*3*512)*512 = 2,359,296
CONV3-512: [28x28x512] memory: 28*28*512=400K params: (3*3*512)*512 = 2,359,296
POOL2: [14x14x512] memory: 14*14*512=100K params: 0
CONV3-512: [14x14x512] memory: 14*14*512=100K params: (3*3*512)*512 = 2,359,296
CONV3-512: [14x14x512] memory: 14*14*512=100K params: (3*3*512)*512 = 2,359,296
CONV3-512: [14x14x512] memory: 14*14*512=100K params: (3*3*512)*512 = 2,359,296
POOL2: [7x7x512] memory: 7*7*512=25K params: 0
FC: [1x1x4096] memory: 4096 params: 7*7*512*4096 = 102,760,448
FC: [1x1x4096] memory: 4096 params: 4096*4096 = 16,777,216
FC: [1x1x1000] memory: 1000 params: 4096*1000 = 4,096,000

TOTAL memory: 24M * 4 bytes ~= 96MB / image (only forward! ~*2 for bwd)
TOTAL params: 138M parameters



VGG16

Common names

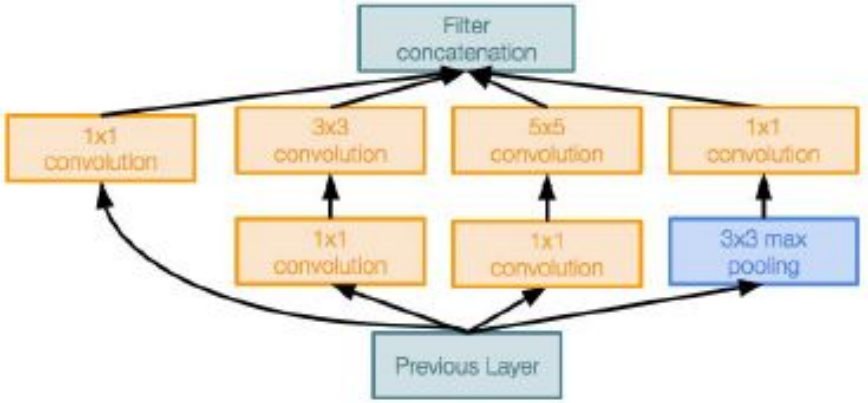
Slide Credit: Fei Fei Li et al., Stanford

Case Study: GoogLeNet

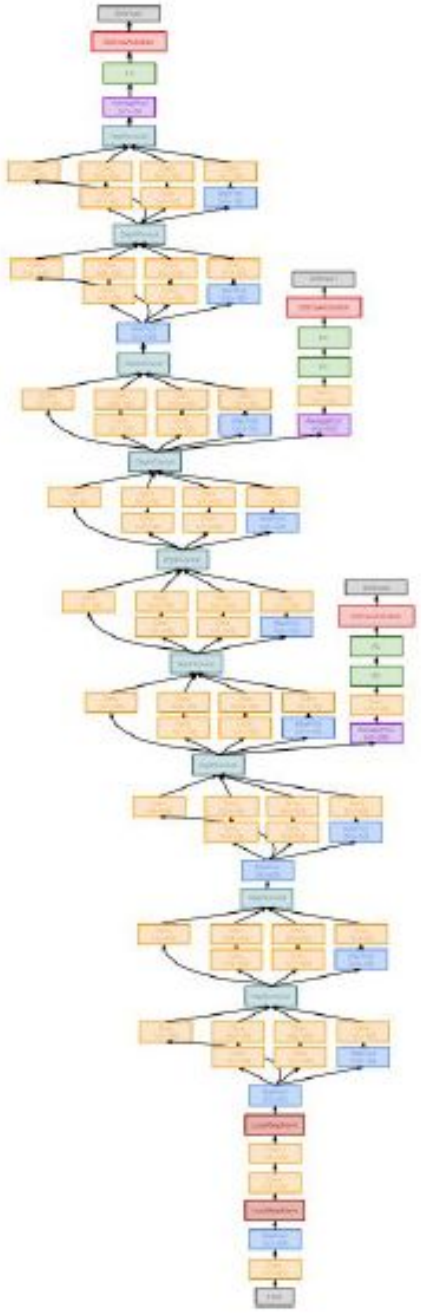
[Szegedy et al., 2014]

Deeper networks, with computational efficiency

- 22 layers
- Efficient "Inception" module
- No FC layers
- Only 5 million parameters!
  12x less than AlexNet
- ILSVRC'14 classification winner
  (6.7% top 5 error)

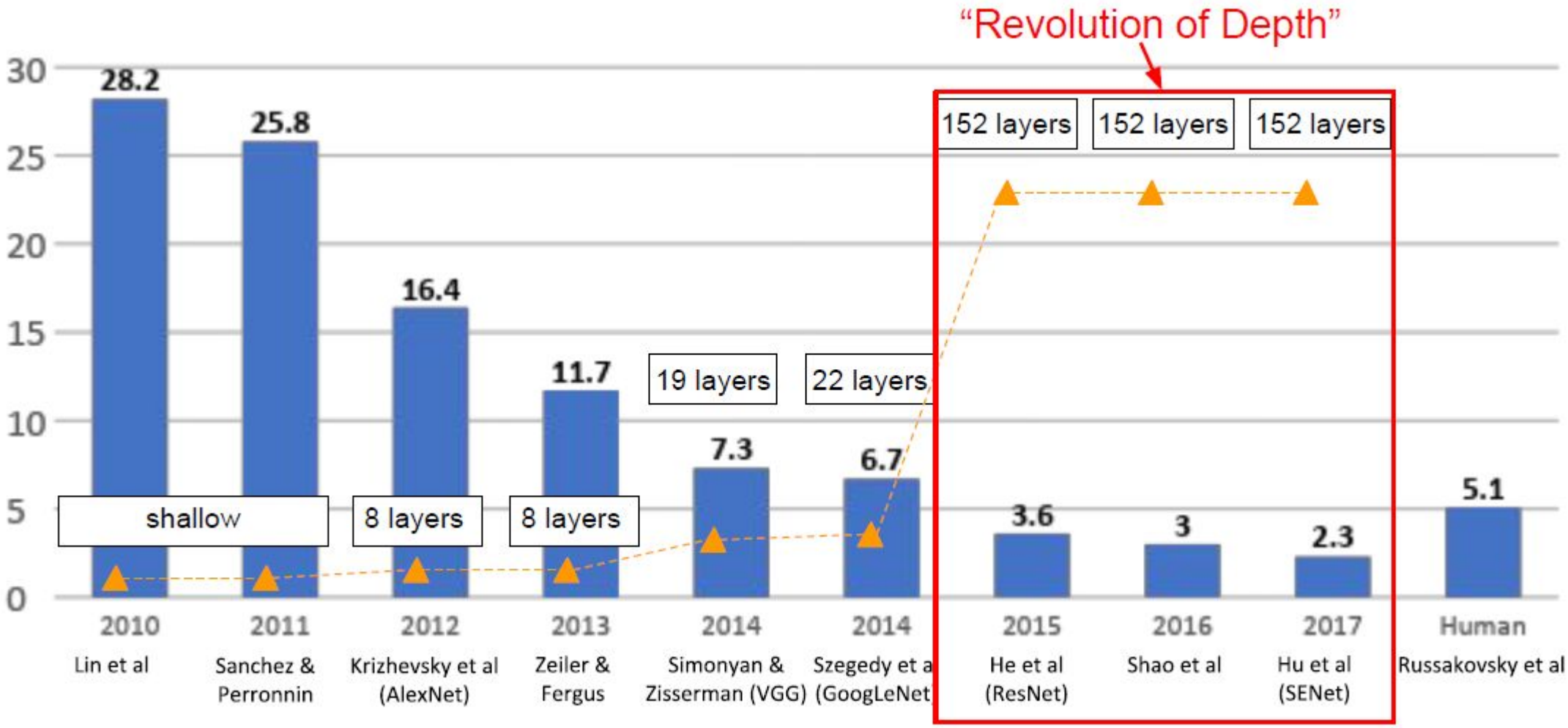


Inception module



Slide Credit: Fei Fei Li et al., Stanford

ImageNet Large Scale Visual Recognition Challenge (ILSVRC) winners



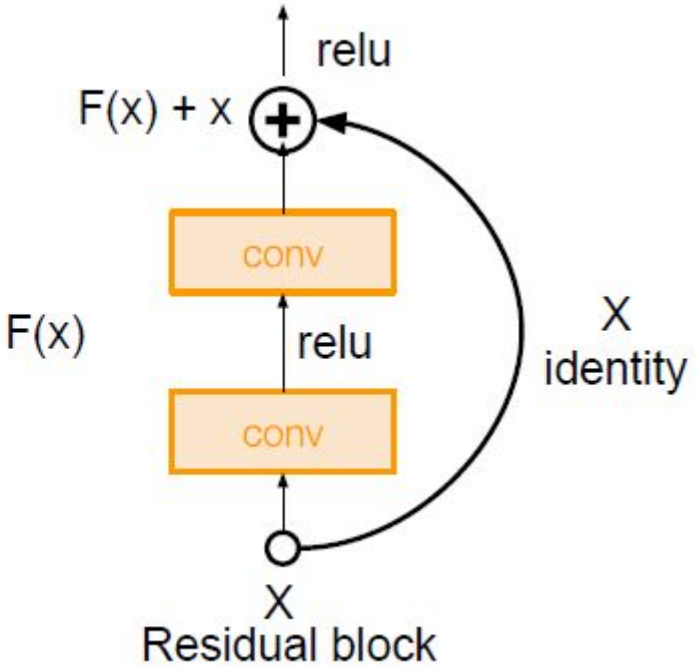
Slide Credit: Fei Fei Li et al., Stanford

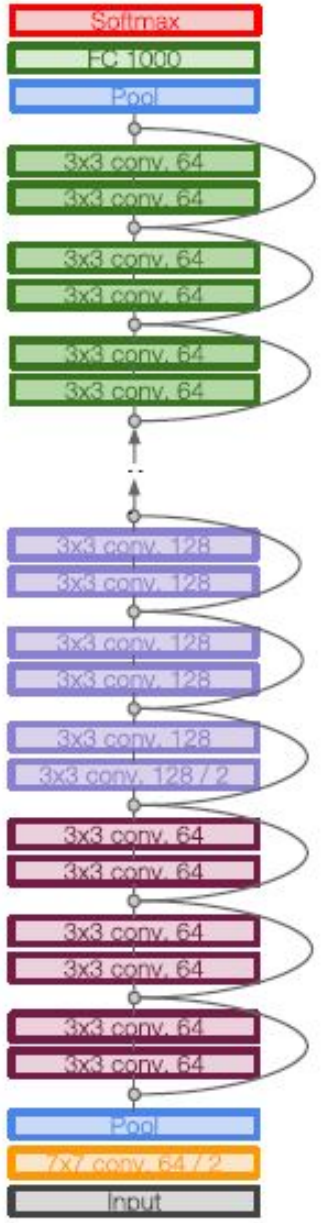
Case Study: ResNet

[He et al., 2015]

Very deep networks using residual connections

- 152-layer model for ImageNet
- ILSVRC'15 classification winner (3.57% top 5 error)
- Swept all classification and detection competitions in ILSVRC'15 and COCO'15!





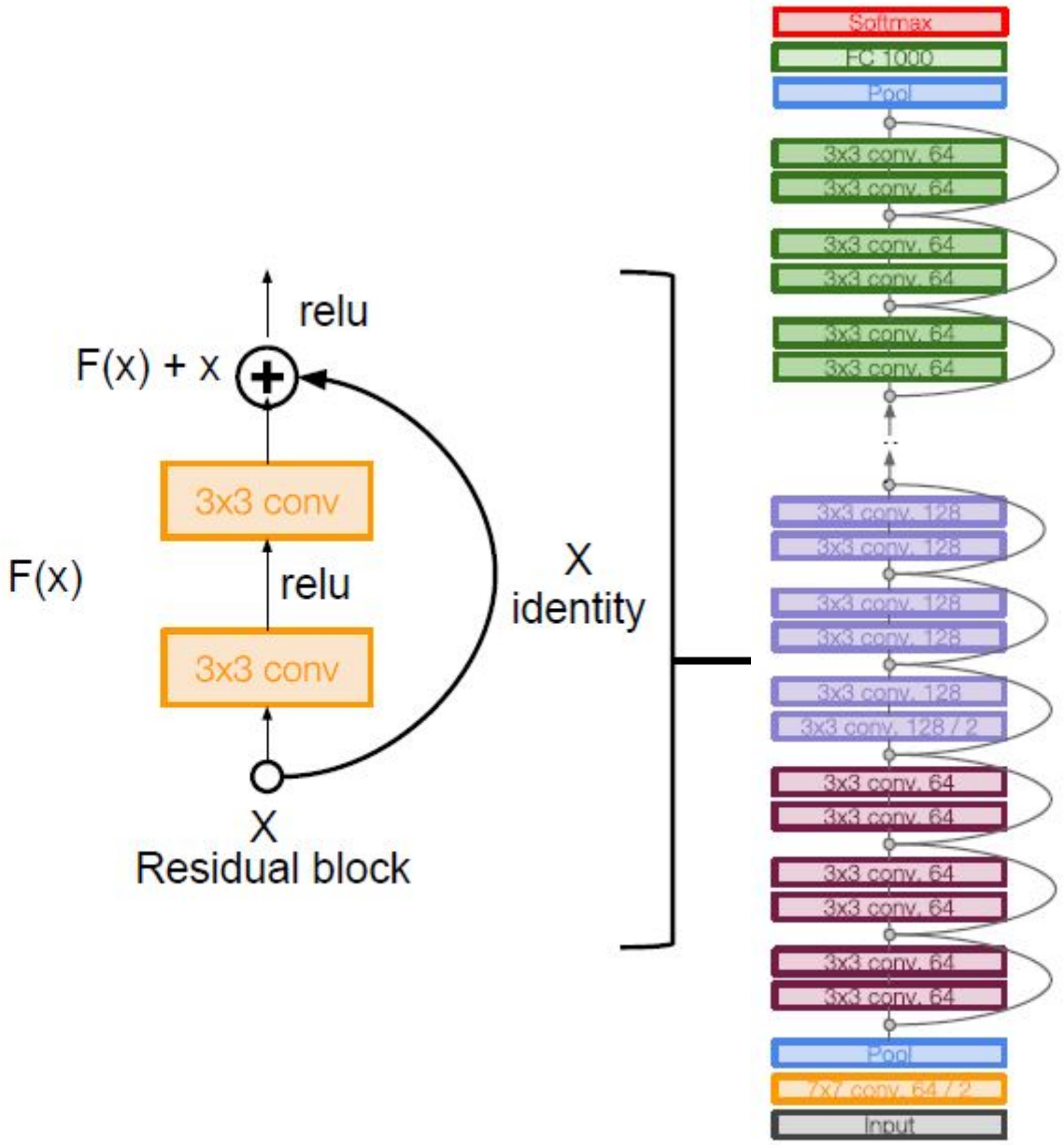
Slide Credit: Fei Fei Li et al., Stanford

Case Study: ResNet

[He et al., 2015]

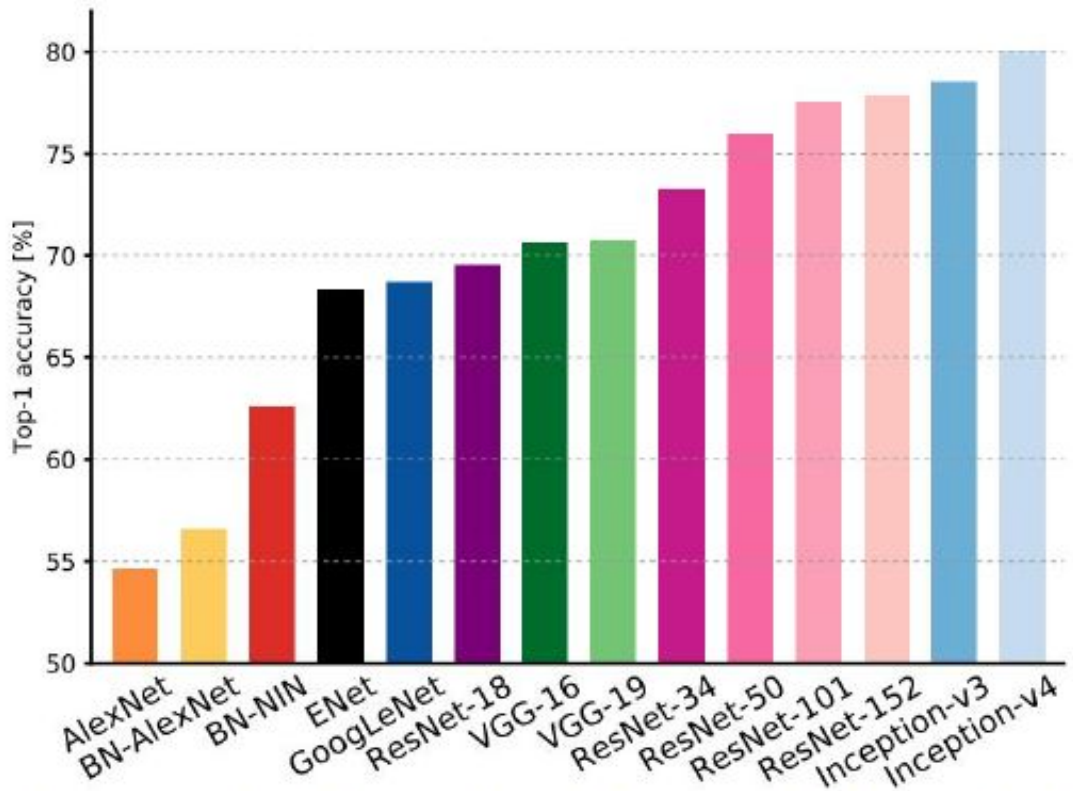
Full ResNet architecture:

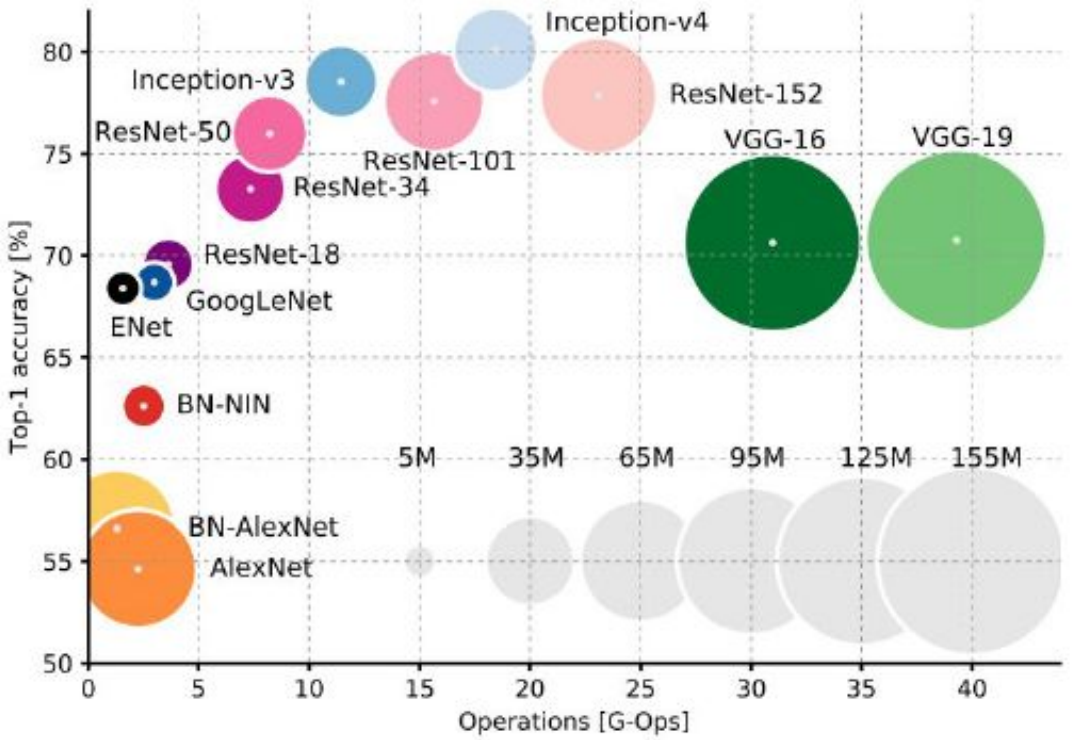
- Stack residual blocks
- Every residual block has two 3x3 conv layers
- Periodically, double # of filters and downsample spatially using stride 2 (/2 in each dimension)
- Additional conv layer at the beginning
- No FC layers at the end (only FC 1000 to output classes)



Slide Credit: Fei Fei Li et al., Stanford

Comparing complexity...



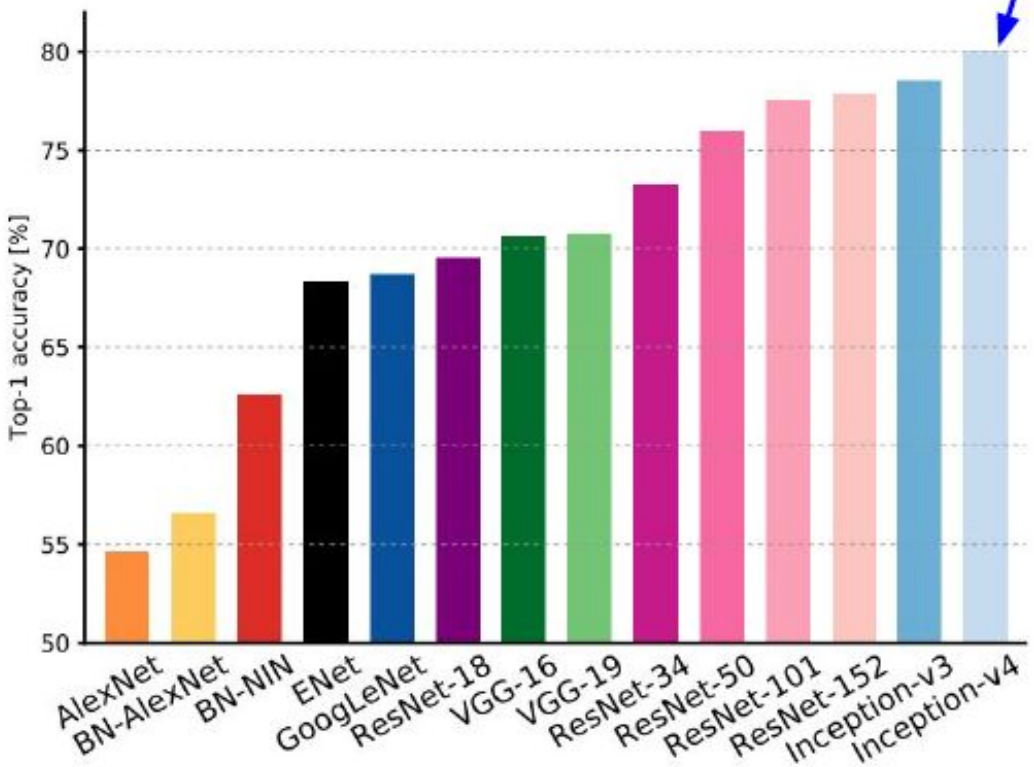


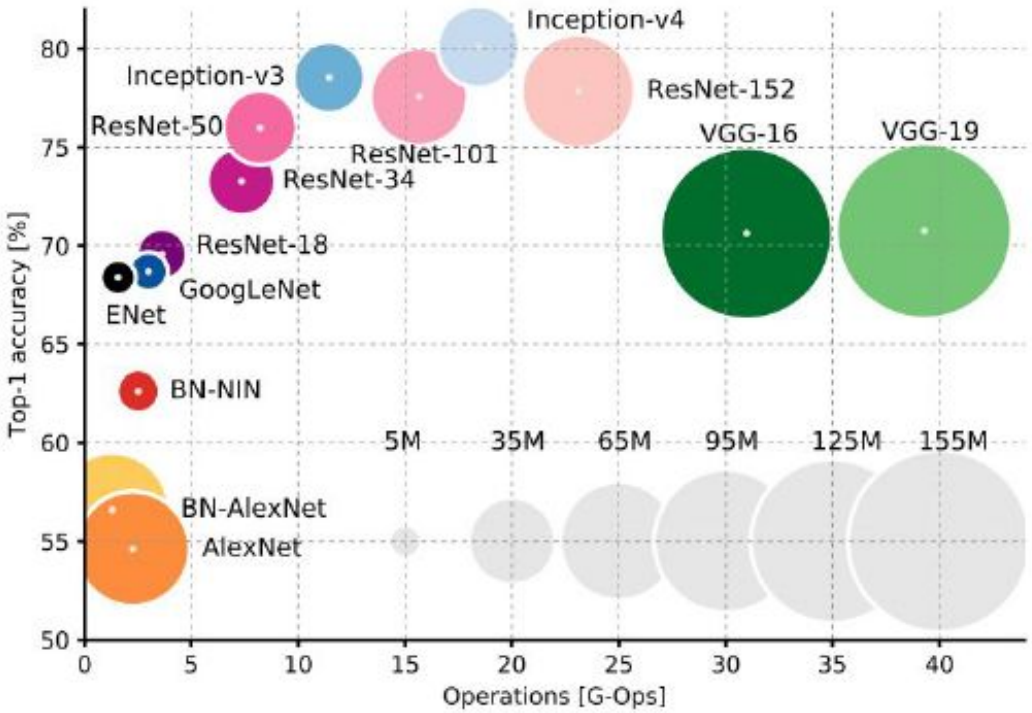
An Analysis of Deep Neural Network Models for Practical Applications, 2017.

Slide Credit: Fei Fei Li et al., Stanford

Comparing complexity...

Inception-v4: Resnet + Inception!

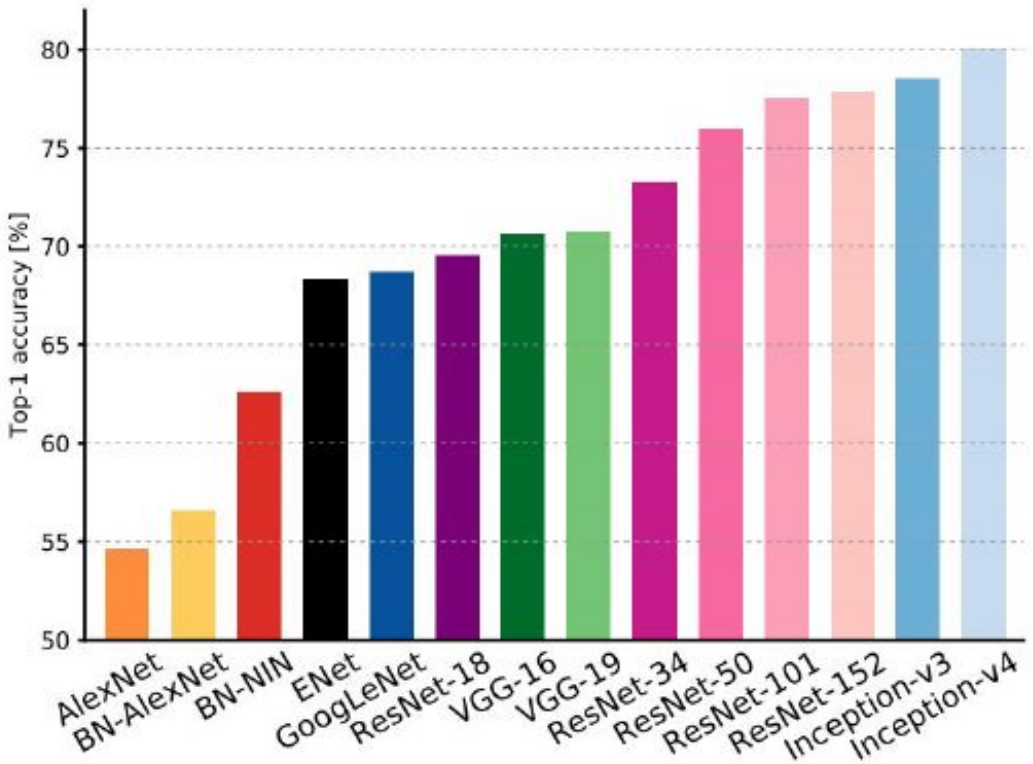




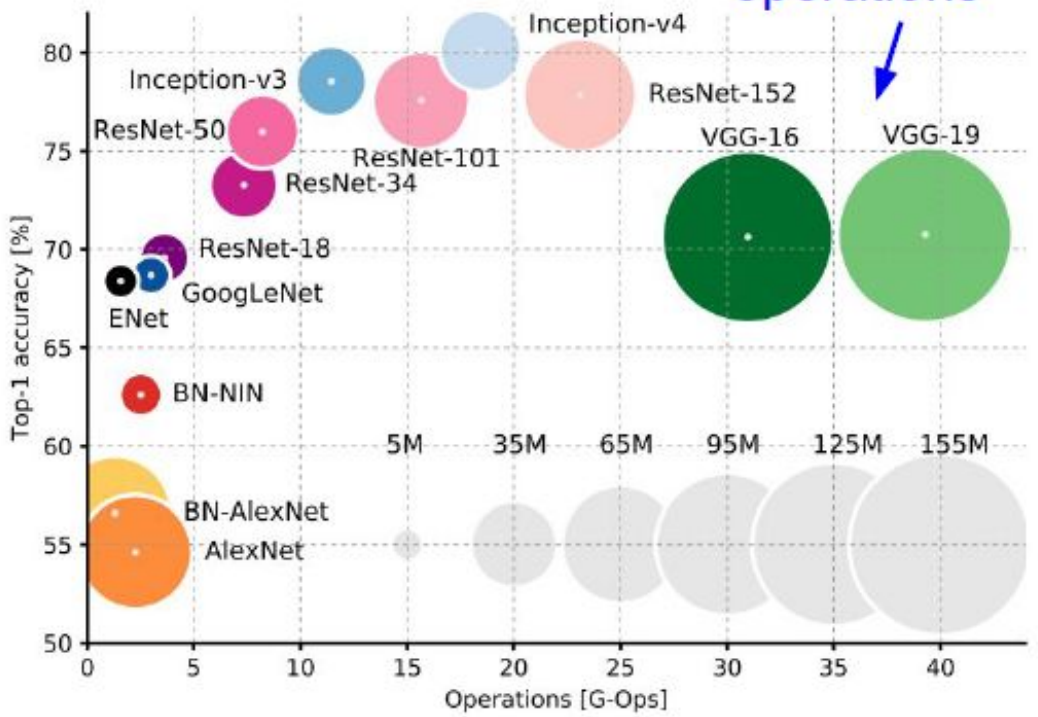
An Analysis of Deep Neural Network Models for Practical Applications, 2017.

Slide Credit: Fei Fei Li et al., Stanford

Comparing complexity...



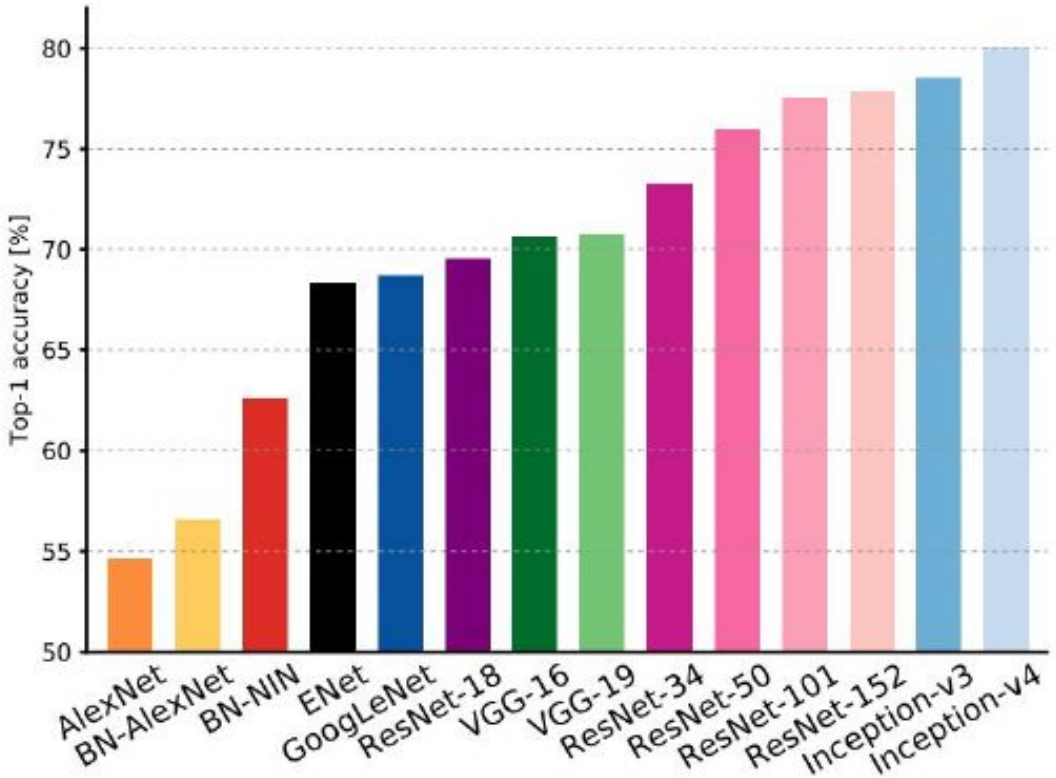
VGG: Highest memory, most operations



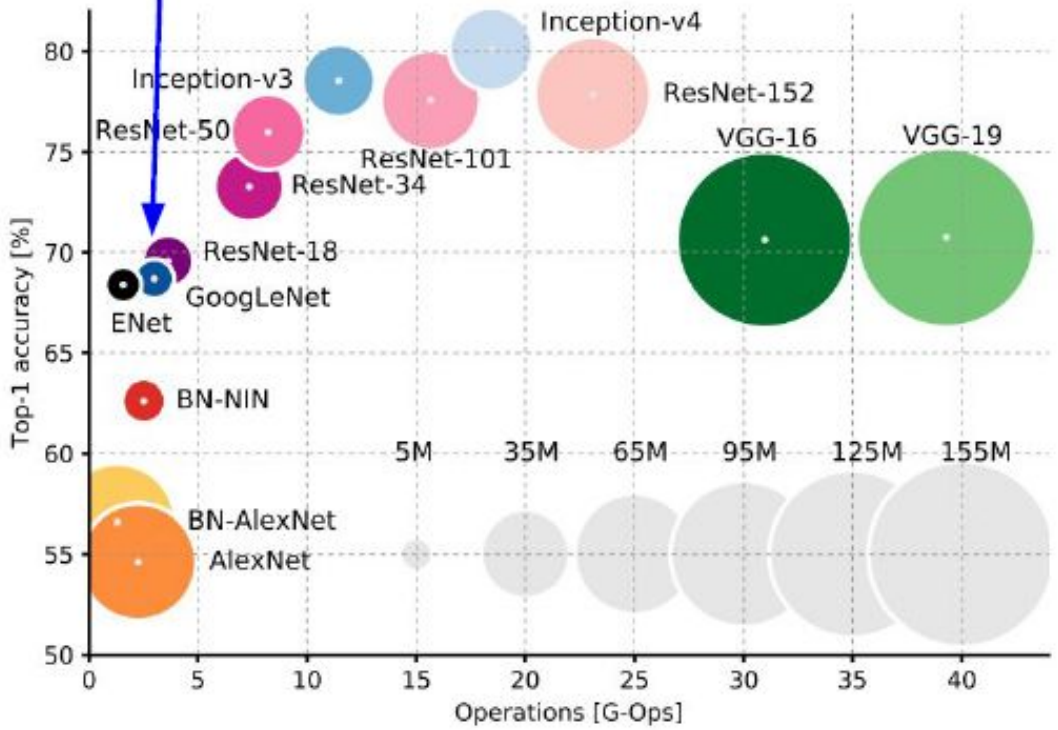
An Analysis of Deep Neural Network Models for Practical Applications, 2017.

Slide Credit: Fei Fei Li et al., Stanford

Comparing complexity...



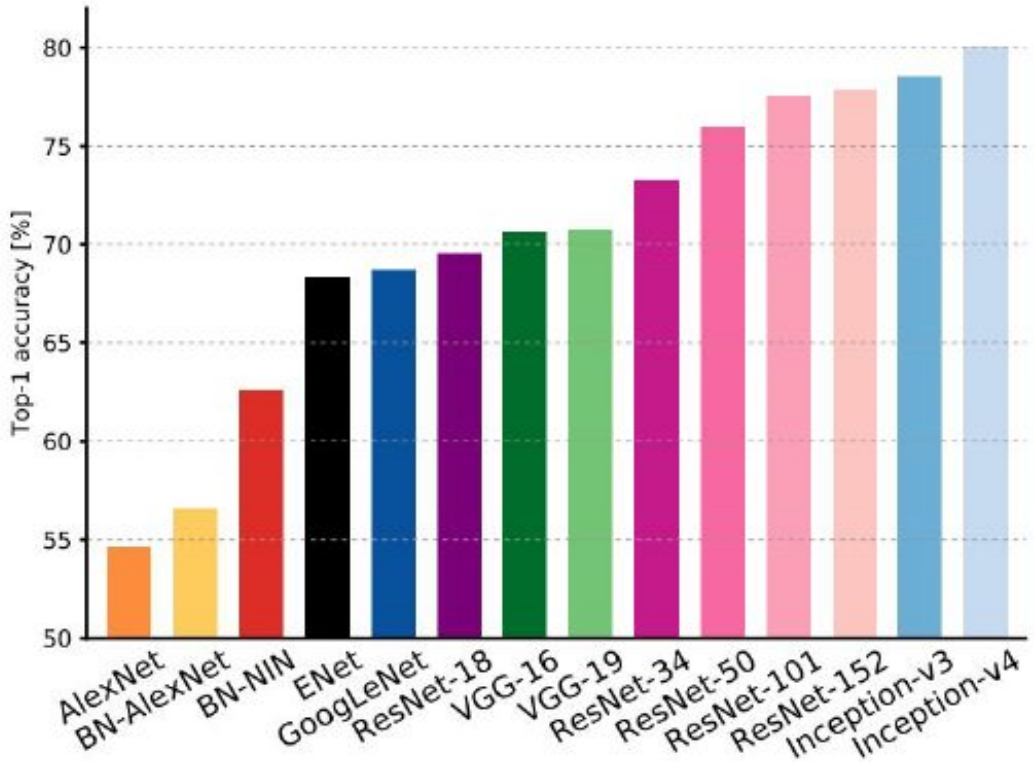
GoogLeNet:
most efficient



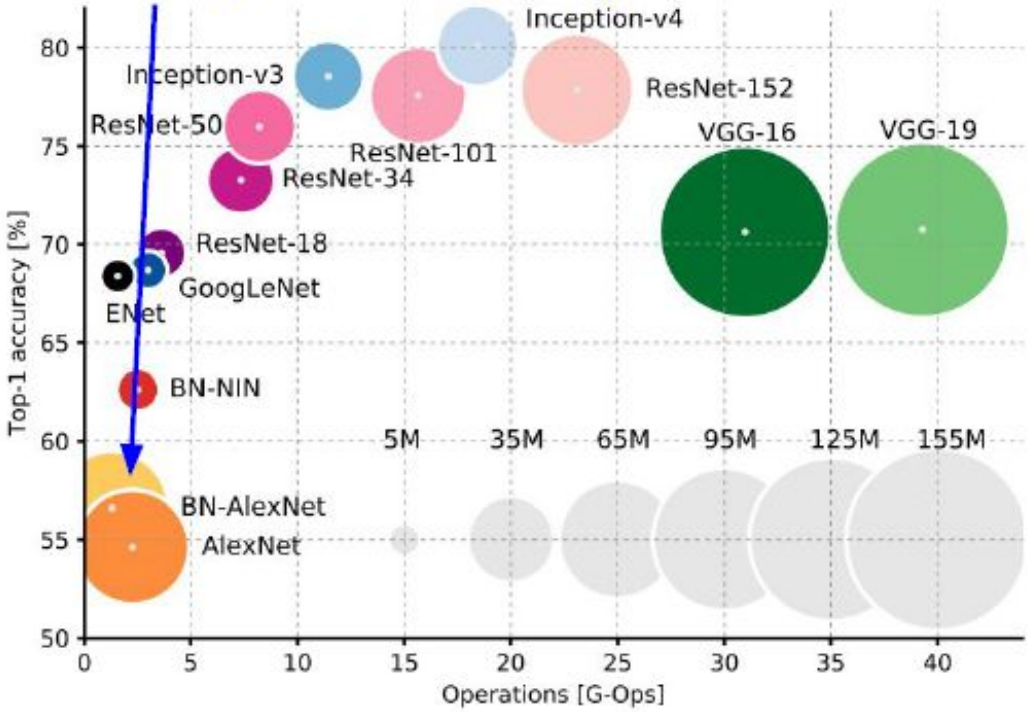
An Analysis of Deep Neural Network Models for Practical Applications, 2017.

Slide Credit: Fei Fei Li et al., Stanford

Comparing complexity...



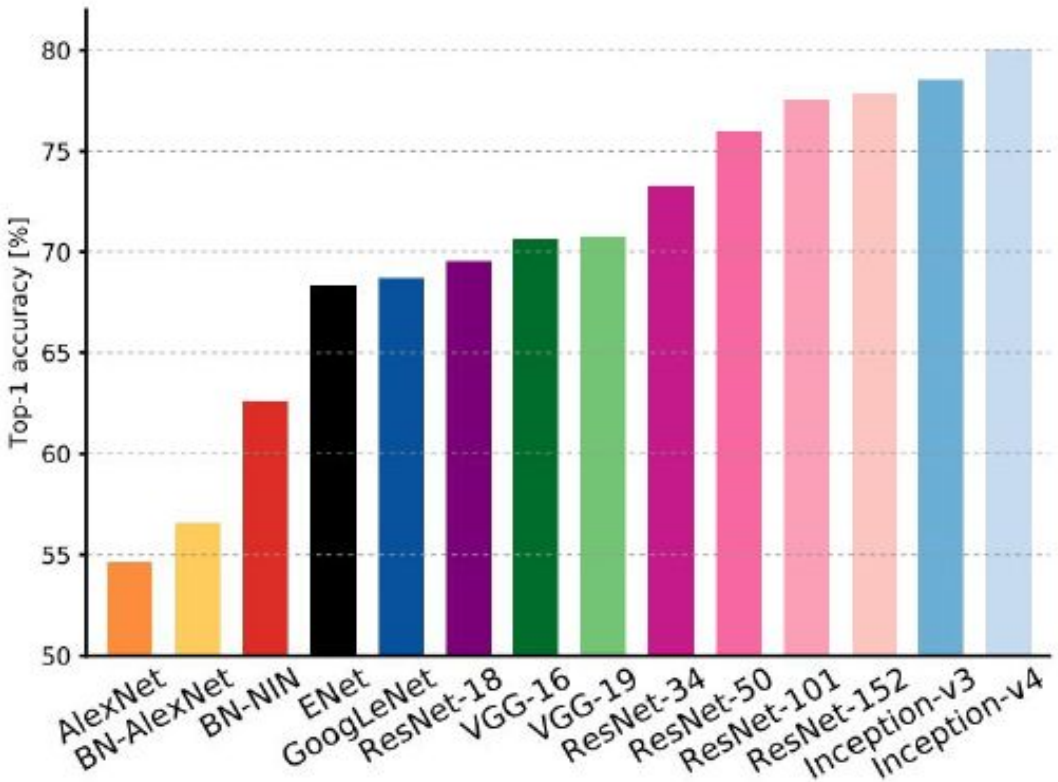
AlexNet:
Smaller compute, still memory heavy, lower accuracy



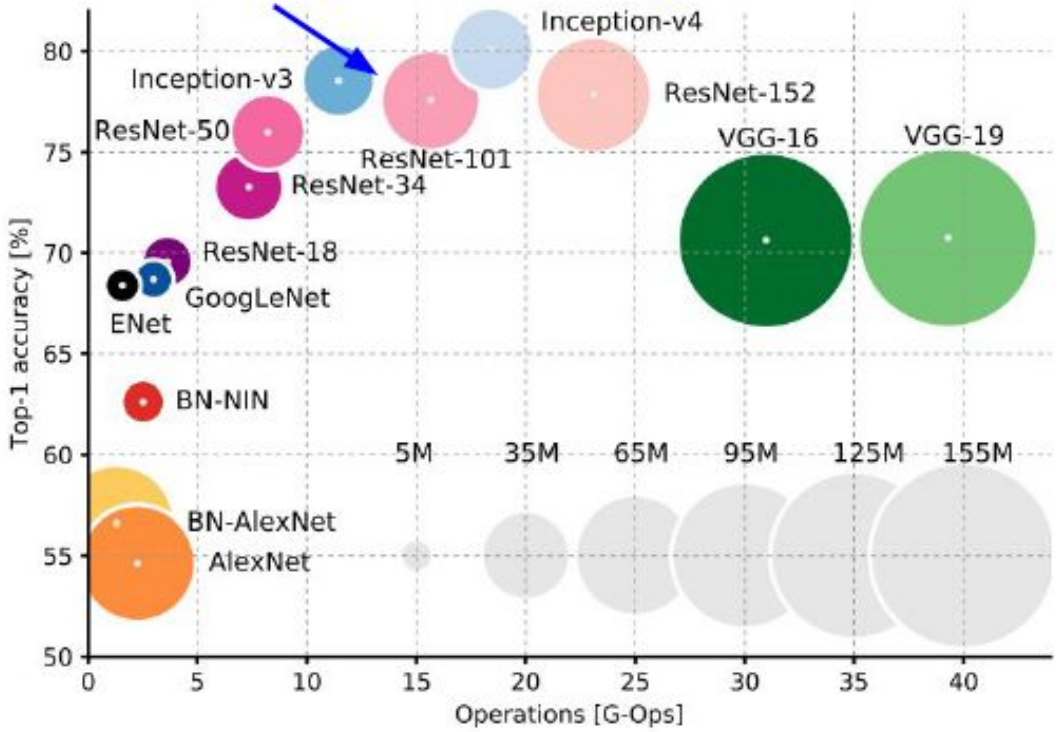
An Analysis of Deep Neural Network Models for Practical Applications, 2017.

Slide Credit: Fei Fei Li et al., Stanford

Comparing complexity...



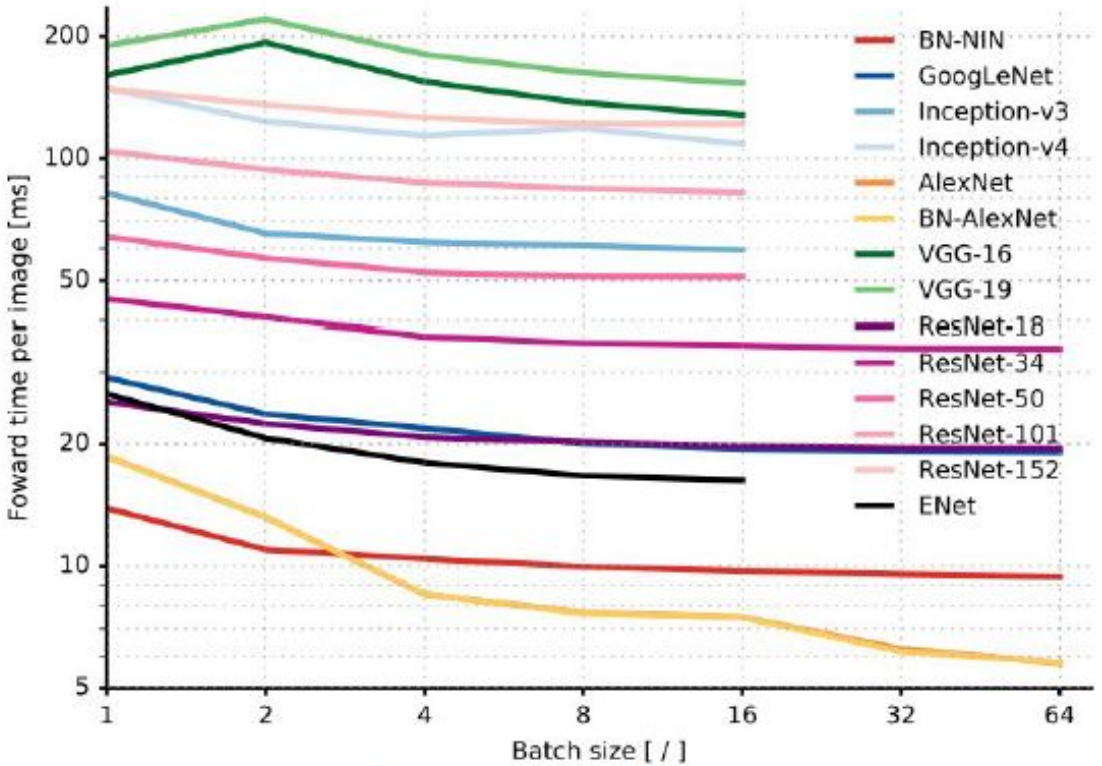
ResNet:
Moderate efficiency depending on model, highest accuracy

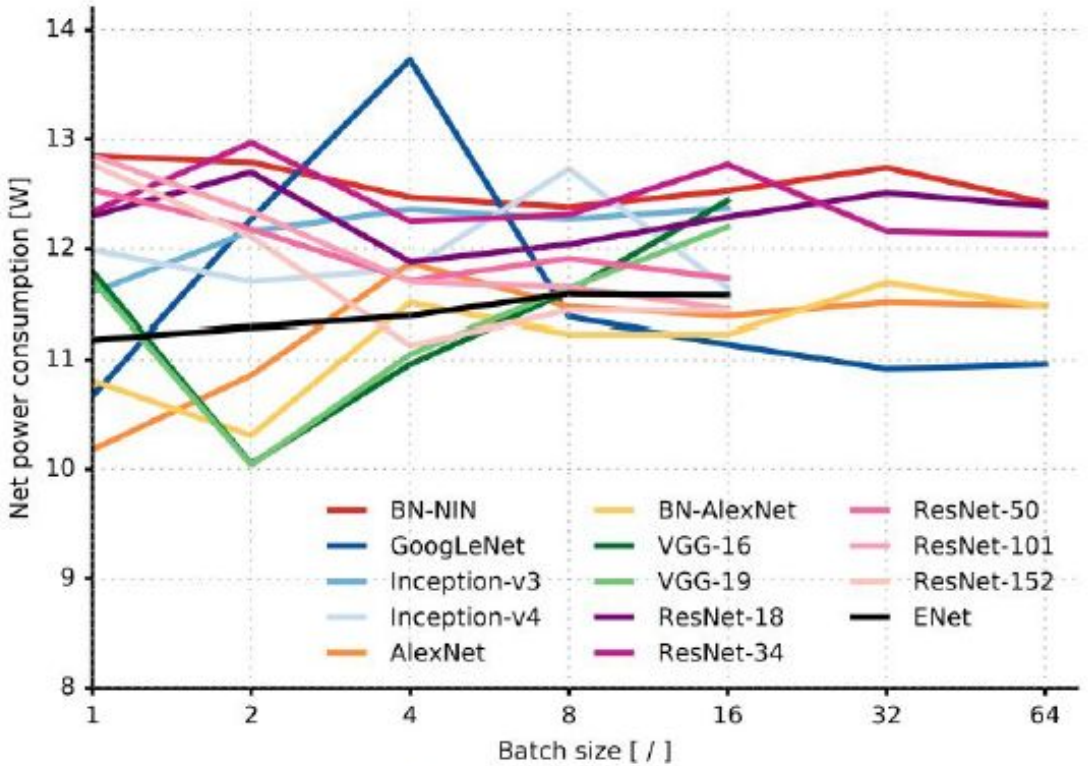


An Analysis of Deep Neural Network Models for Practical Applications, 2017.

Slide Credit: Fei Fei Li et al., Stanford

Forward pass time and power consumption





An Analysis of Deep Neural Network Models for Practical Applications, 2017.

Slide Credit: Fei Fei Li et al., Stanford

Object Detection

Object Detection using Deep Learning

Bikash Santra
National Institutes of Health, USA

Object Detection

Input Image

Question: Where are the cars in the image?

Slide Credit: Kaustav Kundu, University of Toronto

Object Detection

Input Image

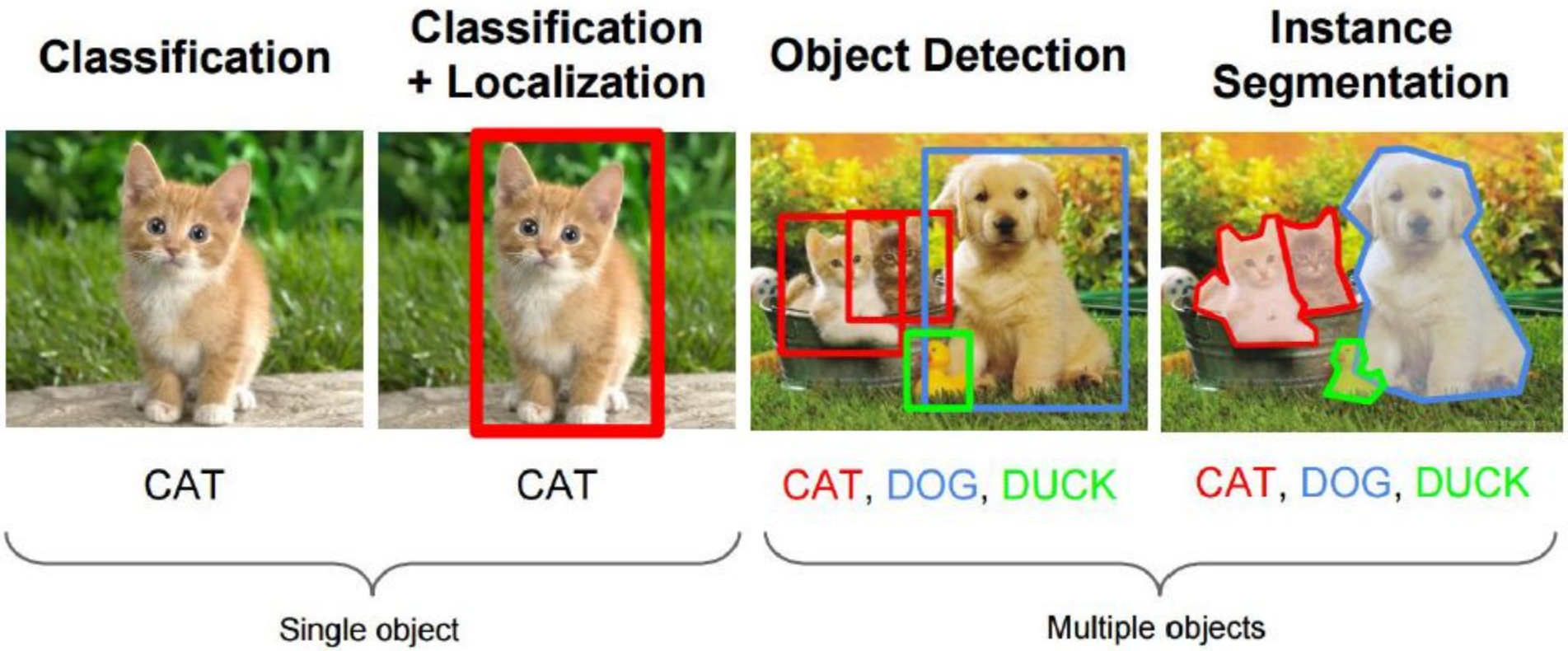
Question: Where are the cars in the image?

Answer:

Object Detection Approach: Recognition + Localization

Slide Credit: Kaustav Kundu, University of Toronto

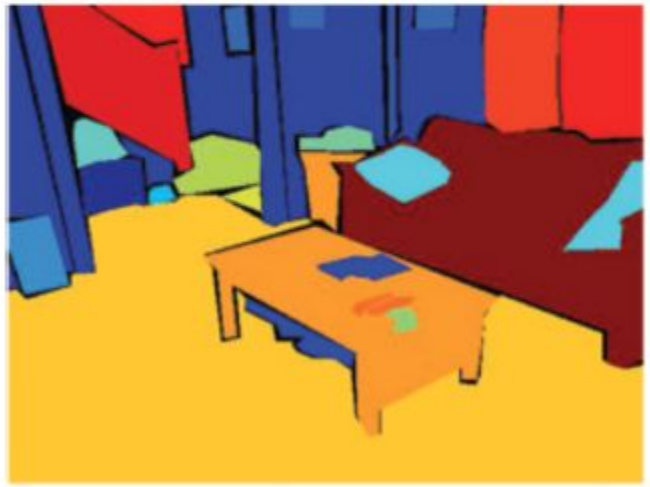
Object Detection



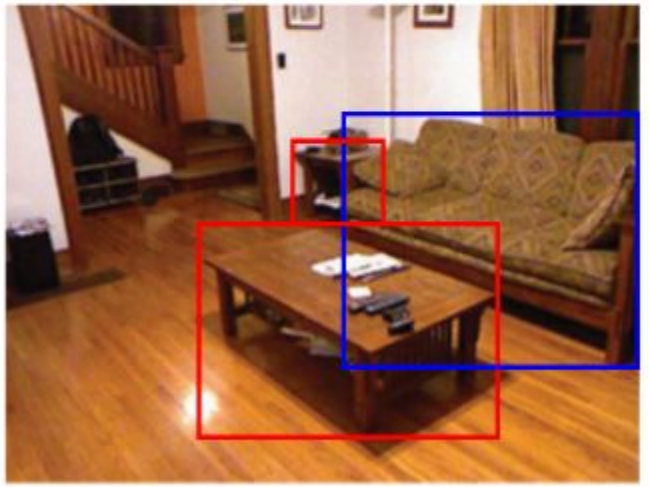
Slide Credit: Sihao Liang et al.

Object Segmentation Vs. Detection

Input Image



Object Segmentation



Object Detection

Slide Credit: Kaustav Kundu, University of Toronto

Typical Object Detection Pipeline

Input Image

$$\begin{bmatrix} \mathbf{x} \end{bmatrix}$$

Feature Extraction

$$f_c(\mathbf{x})$$

Classification

- Candidate Box Selection
- Feature Extraction
- Classification
- Post processing

Slide Credit: Kaustav Kundu, University of Toronto

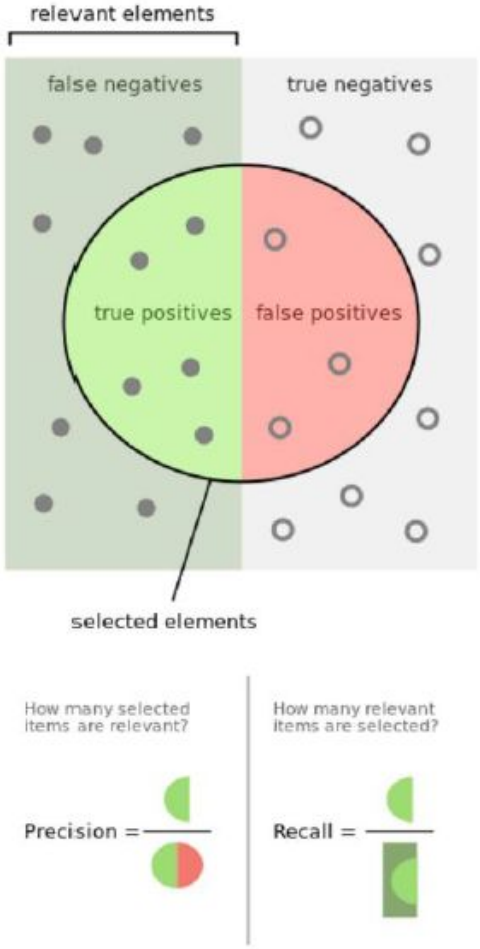
Typical Object Detection Pipeline

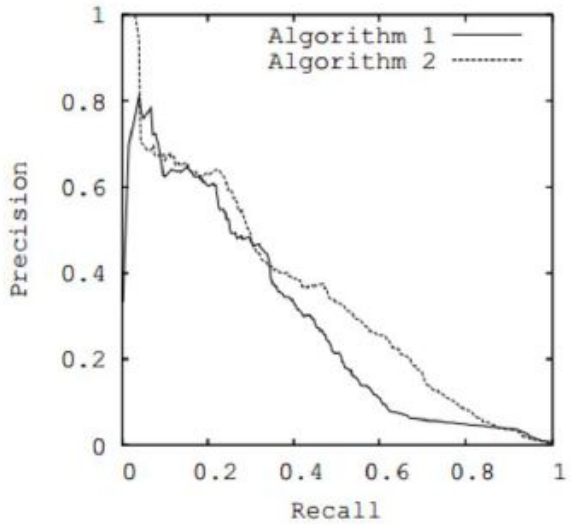
	Candidate Box Selection	Feature Extraction	Classification
Pre Deep Era	Exhaustive	Hand-crafted (*e.g.* HOG)	Linear
RCNN	Region Proposal	Deep	Linear
Fast RCNN	Region Proposal	Deep	
Faster RCNN	Deep	Deep	

Slide Credit: Kaustav Kundu, University of Toronto

Evaluation Indicators for Object Detection

- Recall
- Precision
- mean Average Precision (mAP)
- Intersection over Union (IoU)





Slide Credit: Sihao Liang et al.

Object Detection Competitions

Pascal VOC

COCO

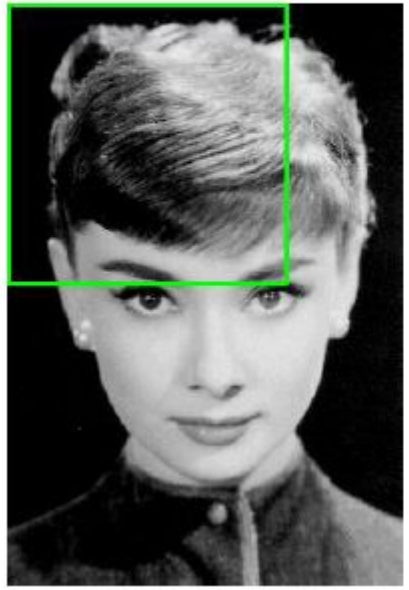
ImageNet ILSVRC

Slide Credit: Sihao Liang et al.

Object Detection: Pre Deep Era

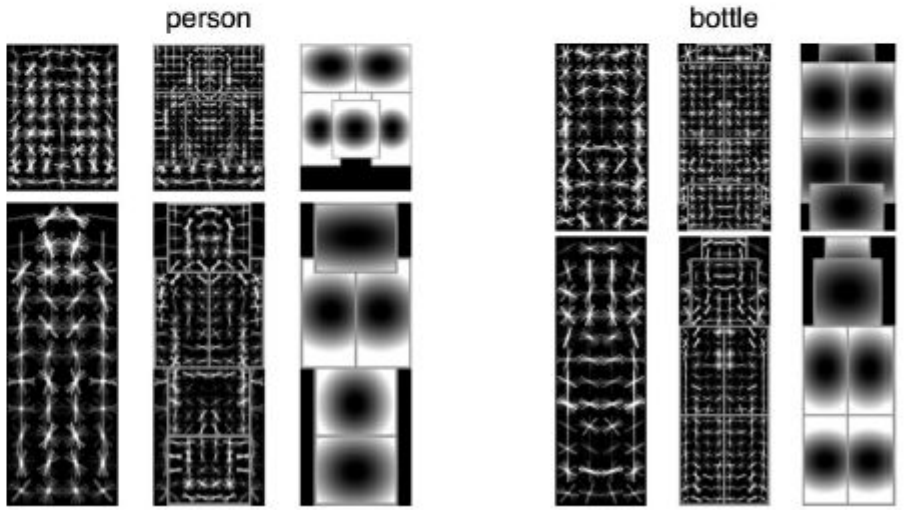
Sliding windows.

- Score every subwindow.



Deformable part models (DPM)

- Uses HOG features
- Very fast



Slide Credit: Sihao Liang et al.

Deep Object Detection

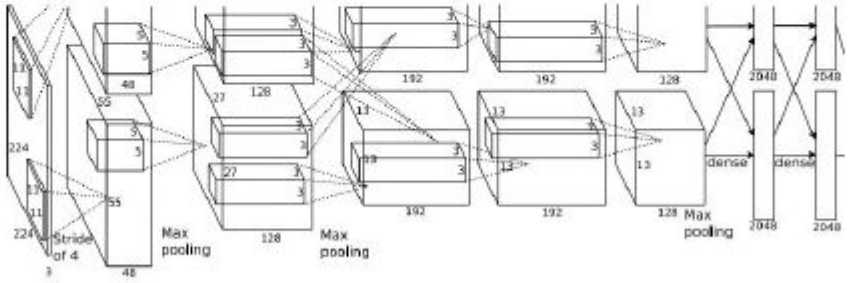
1. Two Stage Detections
2. Unified Detections

Slide Credit: Sihao Liang et al.

Two Stage Detections

Object Detection as Classification: Sliding Window

Apply a CNN to many different crops of the image, CNN classifies each crop as object or background

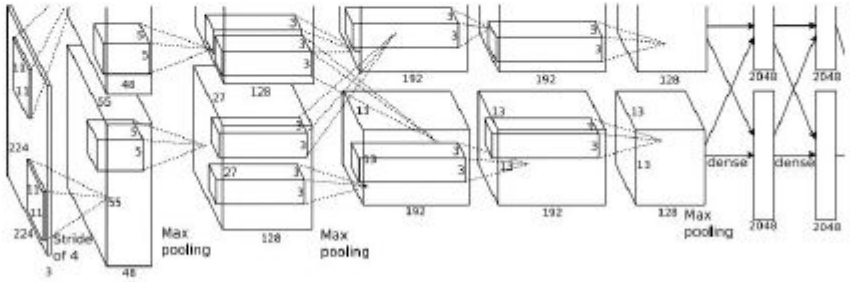


Dog? NO
Cat? NO
Background? YES

Slide Credit: Fei Fei Li et al. Stanford

Object Detection as Classification: Sliding Window

Apply a CNN to many different crops of the image, CNN classifies each crop as object or background

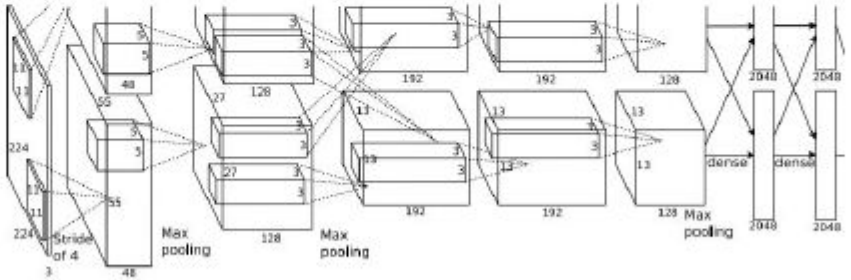


Dog? YES
Cat? NO
Background? NO

Slide Credit: Fei Fei Li et al. Stanford

Object Detection as Classification: Sliding Window

Apply a CNN to many different crops of the image, CNN classifies each crop as object or background

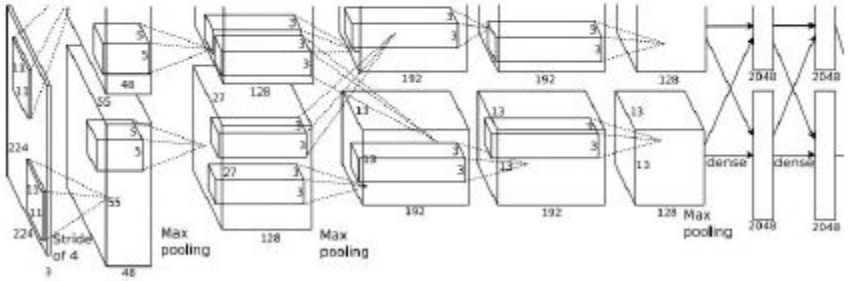


Dog? YES
Cat? NO
Background? NO

Slide Credit: Fei Fei Li et al. Stanford

Object Detection as Classification: Sliding Window

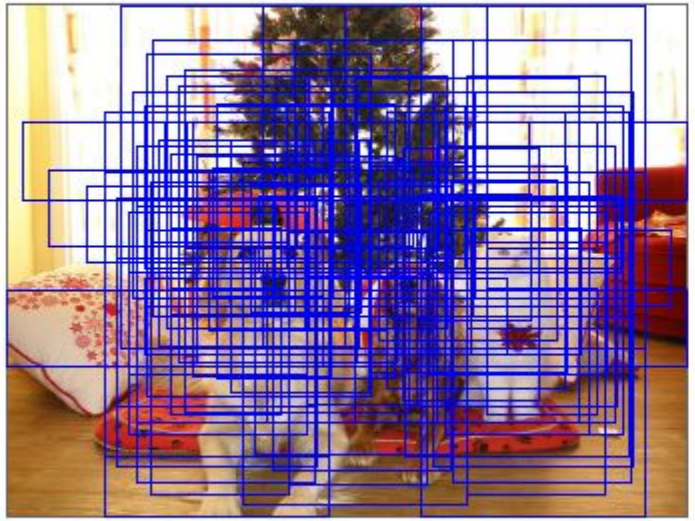
Apply a CNN to many different crops of the image, CNN classifies each crop as object or background



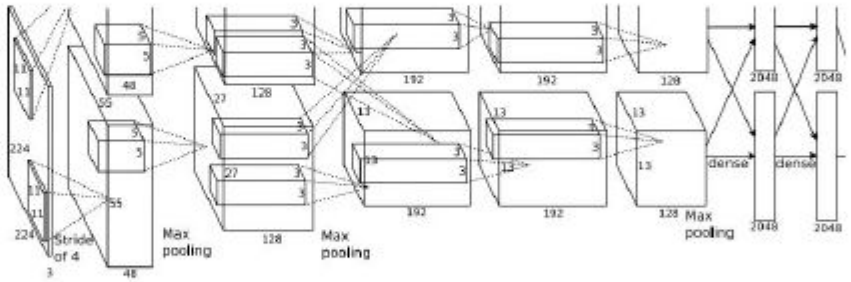
Dog? NO
Cat? YES
Background? NO

Slide Credit: Fei Fei Li et al. Stanford

Object Detection as Classification: Sliding Window



Apply a CNN to many different crops of the image, CNN classifies each crop as object or background



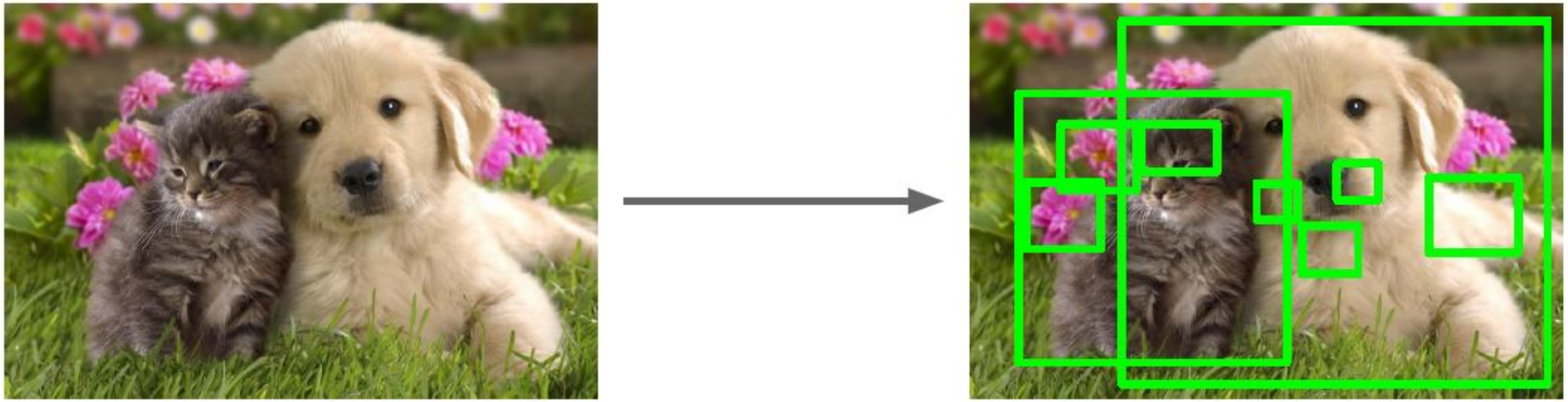
Dog? NO
Cat? YES
Background? NO

Problem: Need to apply CNN to huge number of locations, scales, and aspect ratios, very computationally expensive!

Slide Credit: Fei Fei Li et al. Stanford

Region Proposals / Selective Search

- Find “blobby” image regions that are likely to contain objects
- Relatively fast to run; e.g. Selective Search gives 2000 region proposals in a few seconds on CPU



Alexe et al, "Measuring the objectness of image windows", TPAMI 2012
Uijlings et al, "Selective Search for Object Recognition", IJCV 2013
Cheng et al, "BING: Binarized normed gradients for objectness estimation at 300fps", CVPR 2014
Zitnick and Dollar, "Edge boxes: Locating object proposals from edges", ECCV 2014

Slide Credit: Fei Fei Li et al. Stanford

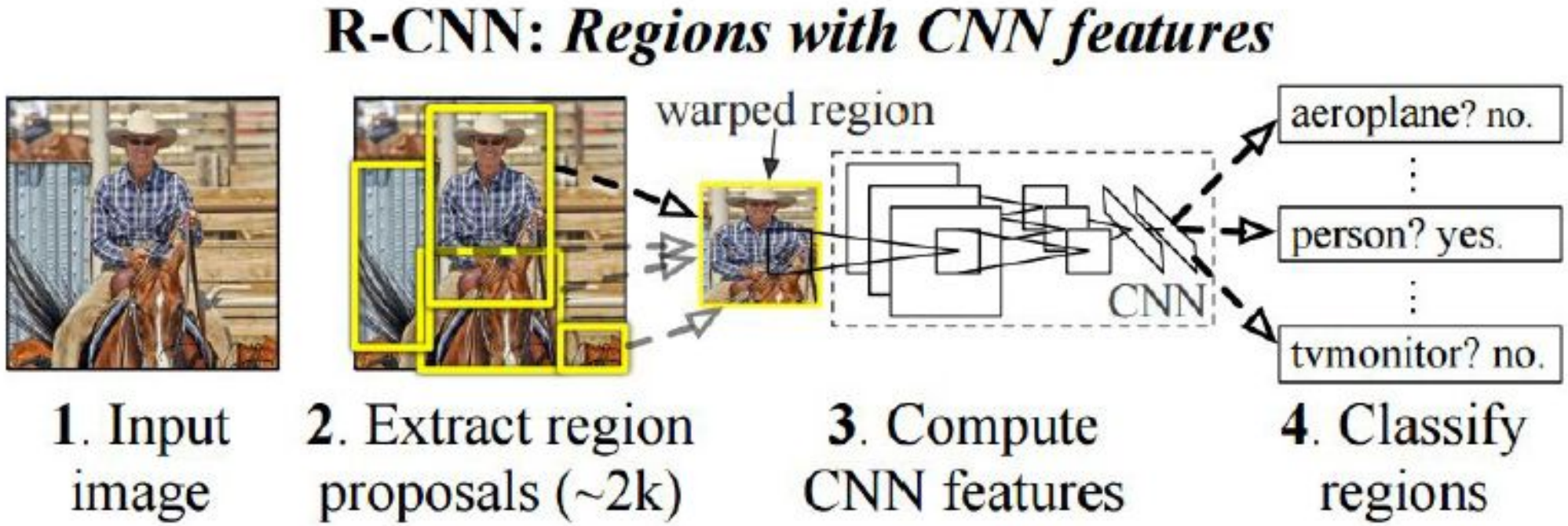
RCNN : Region Proposal + CNN

1. Use selective search to come up with regional proposal
2. First object detection method using CNN

Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.

Slide Credit: Sihao Liang et al.

RCNN



Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.

Slide Credit: Sihao Liang et al.

RCNN

Input image

Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

Slide Credit: Fei Fei Li et al. Stanford

RCNN

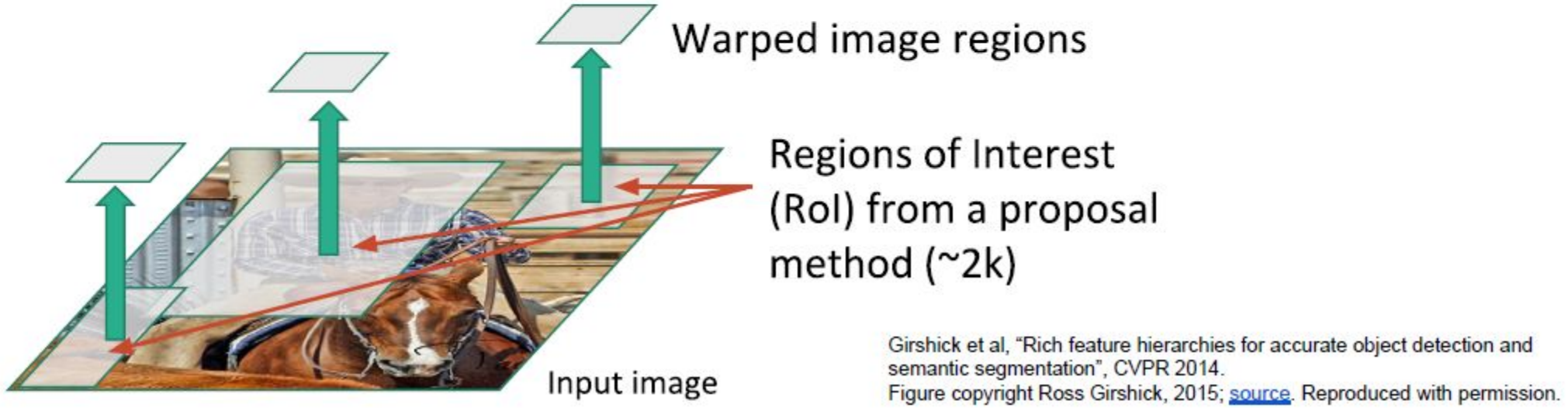
Input image

Regions of Interest (RoI) from a proposal method (~2k)

Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

Slide Credit: Fei Fei Li et al. Stanford

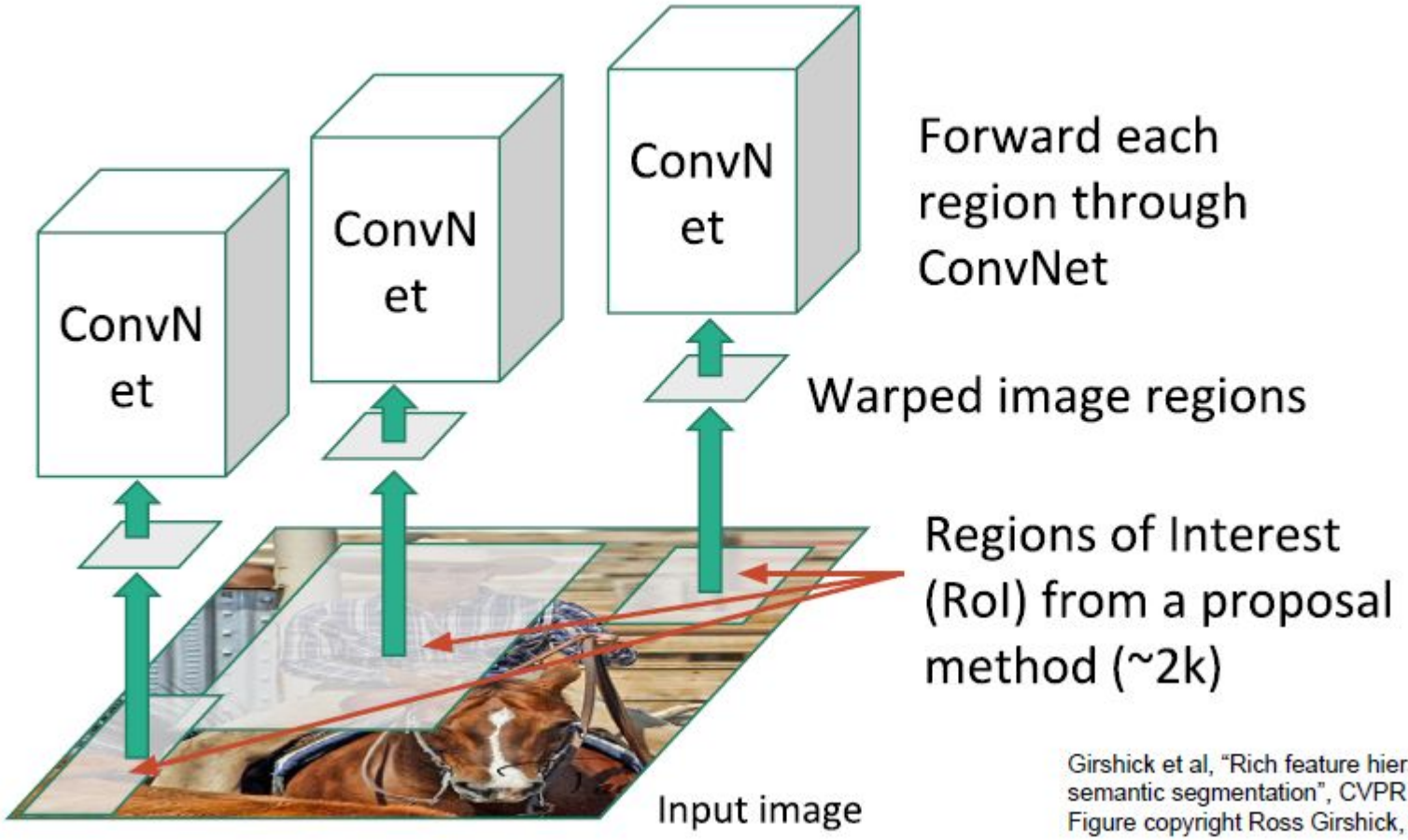
RCNN



Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

Slide Credit: Fei Fei Li et al. Stanford

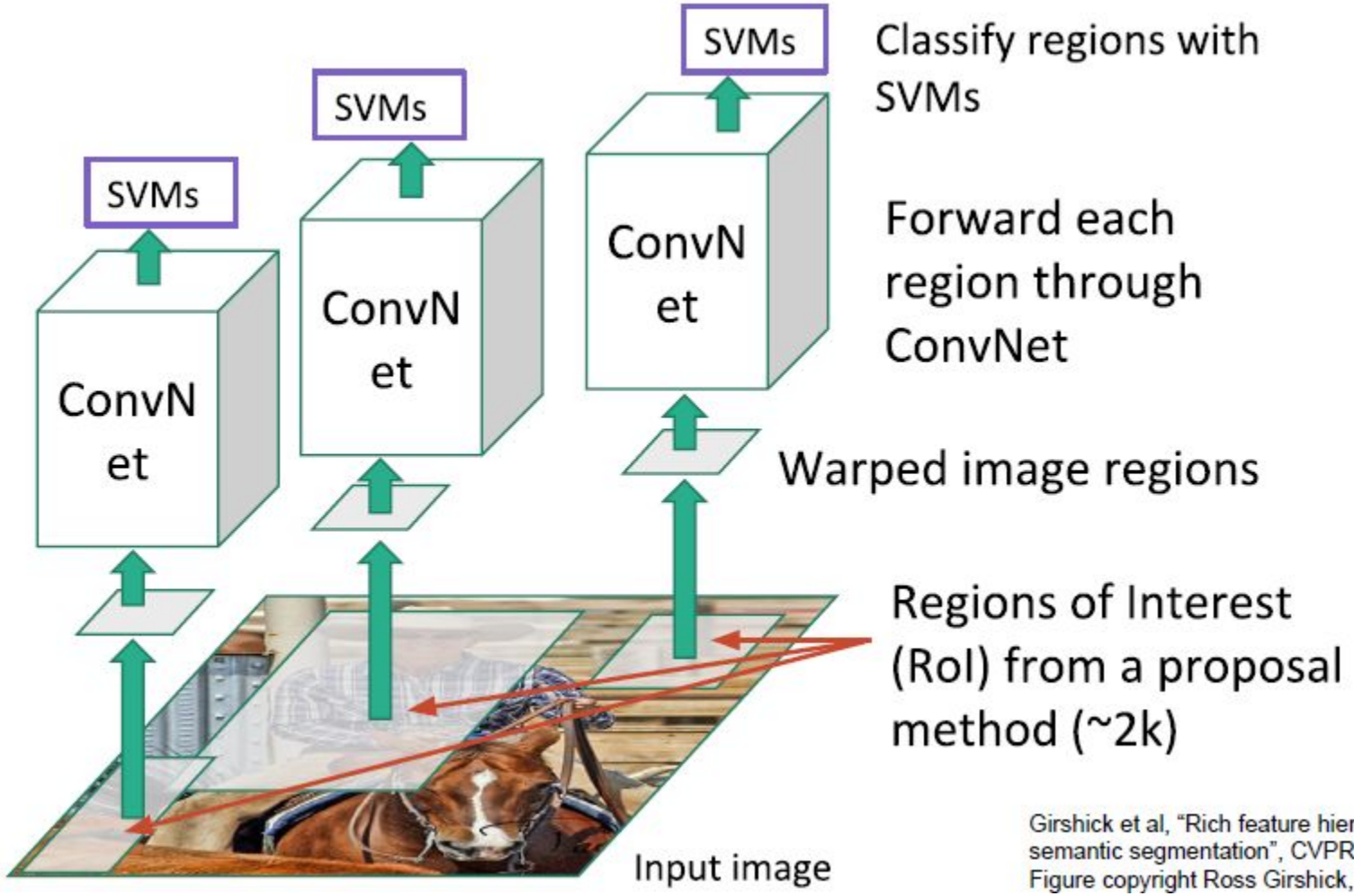
RCNN



Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

Slide Credit: Fei Fei Li et al. Stanford

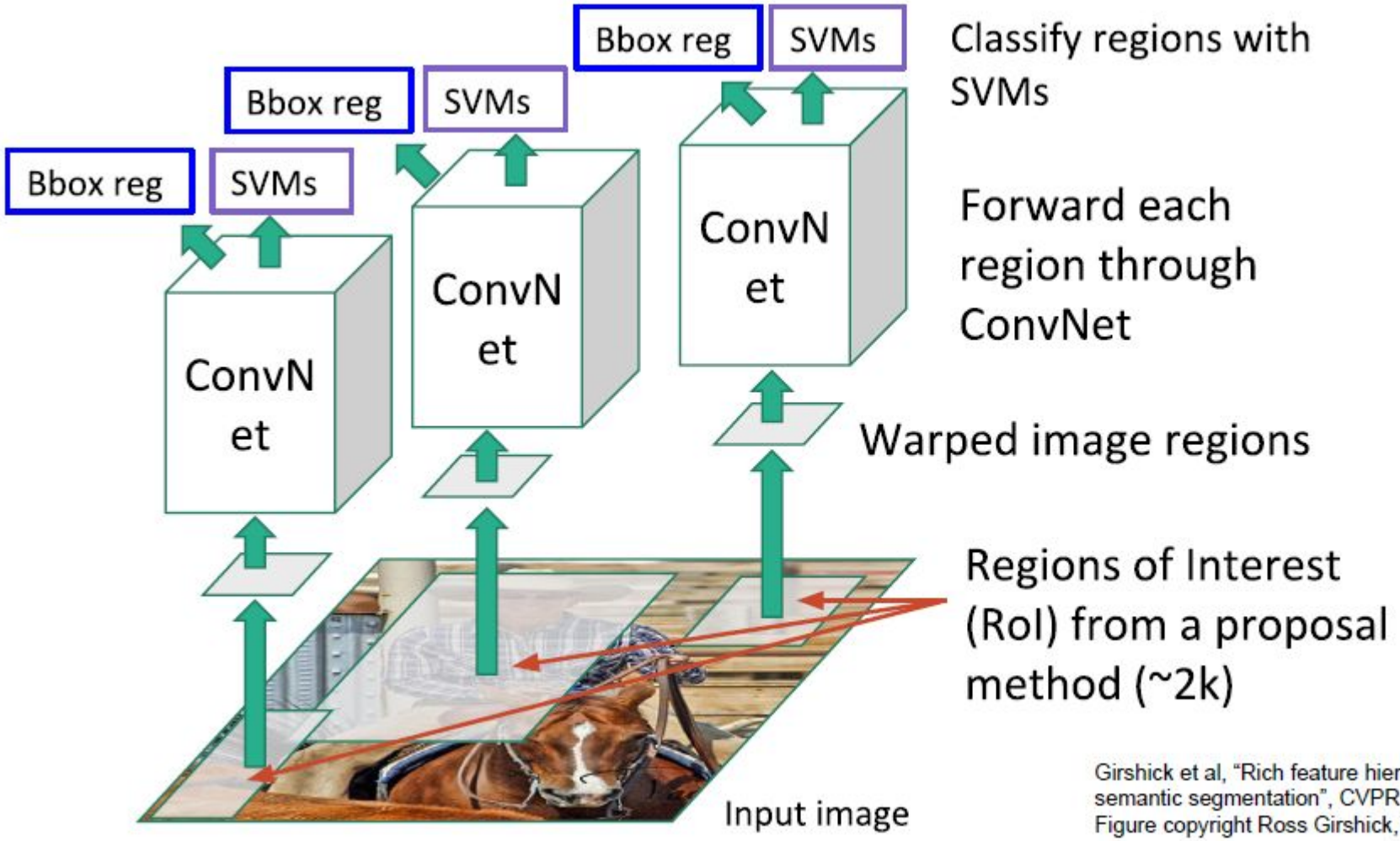
RCNN



Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

Slide Credit: Fei Fei Li et al. Stanford

RCNN

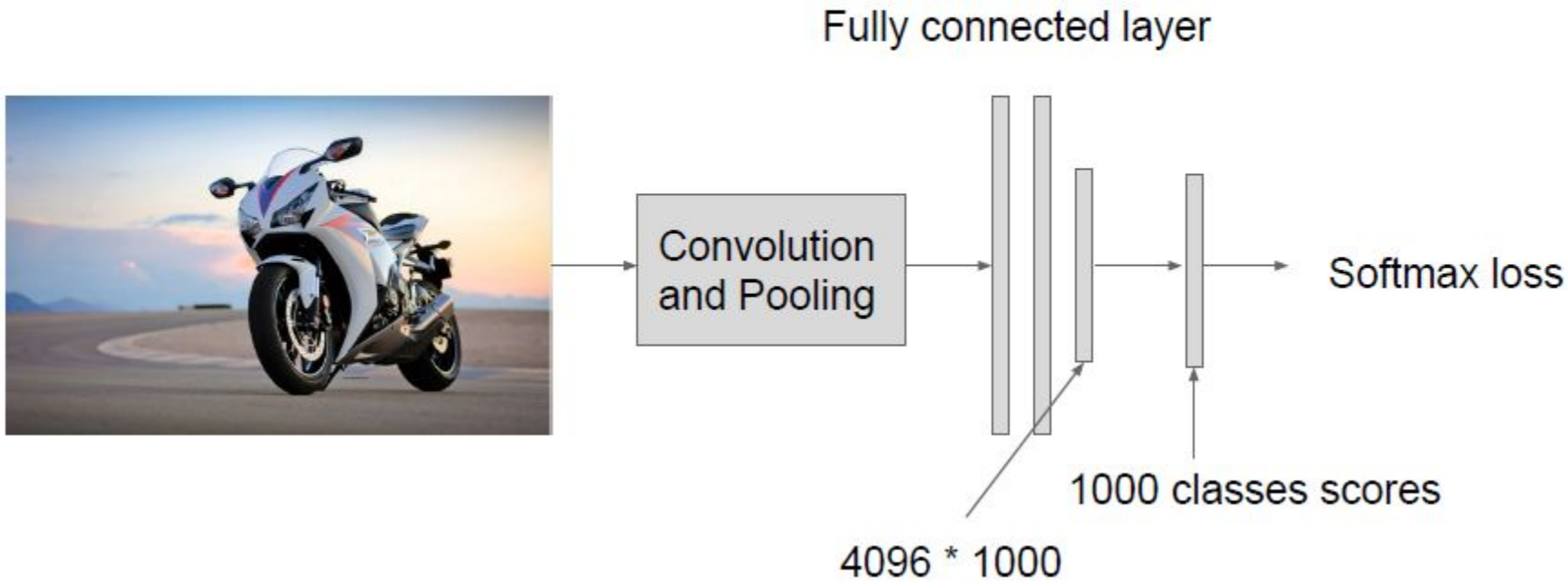


Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

Slide Credit: Fei Fei Li et al. Stanford

Training RCNN

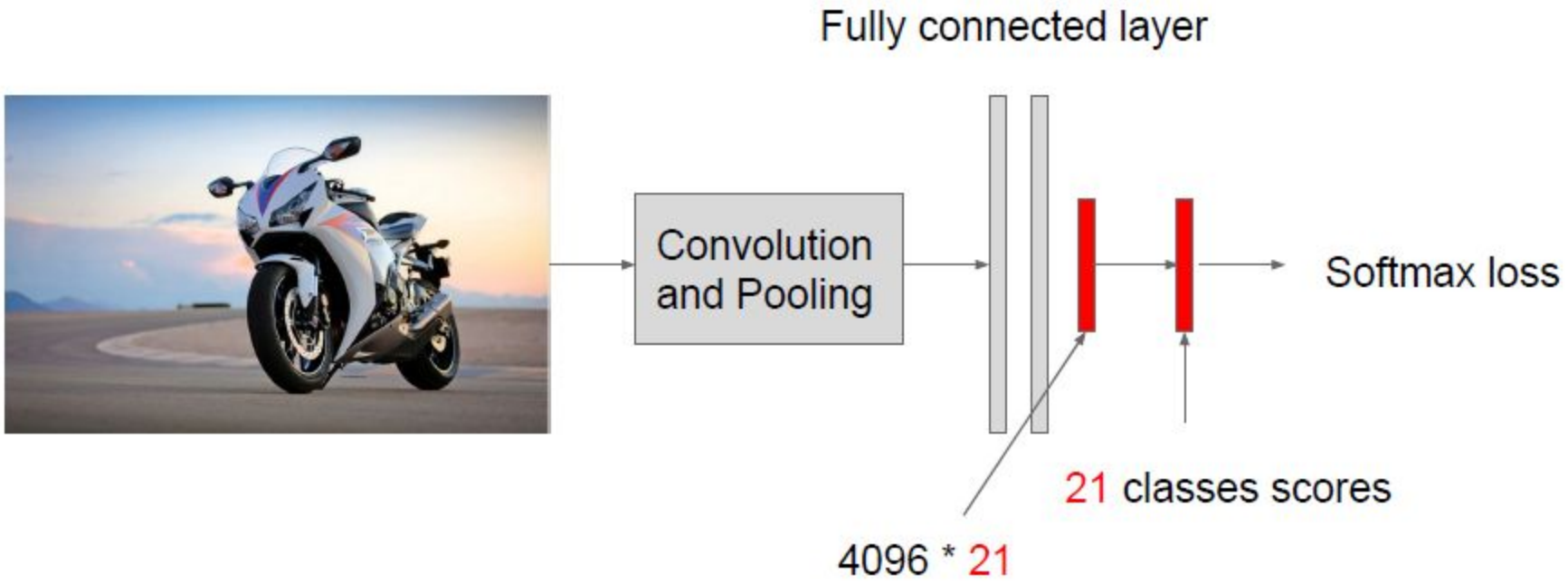
Step1: train your own CNN model for classification (or use existing model), using ImageNet dataset.



Slide Credit: Sihao Liang et al.

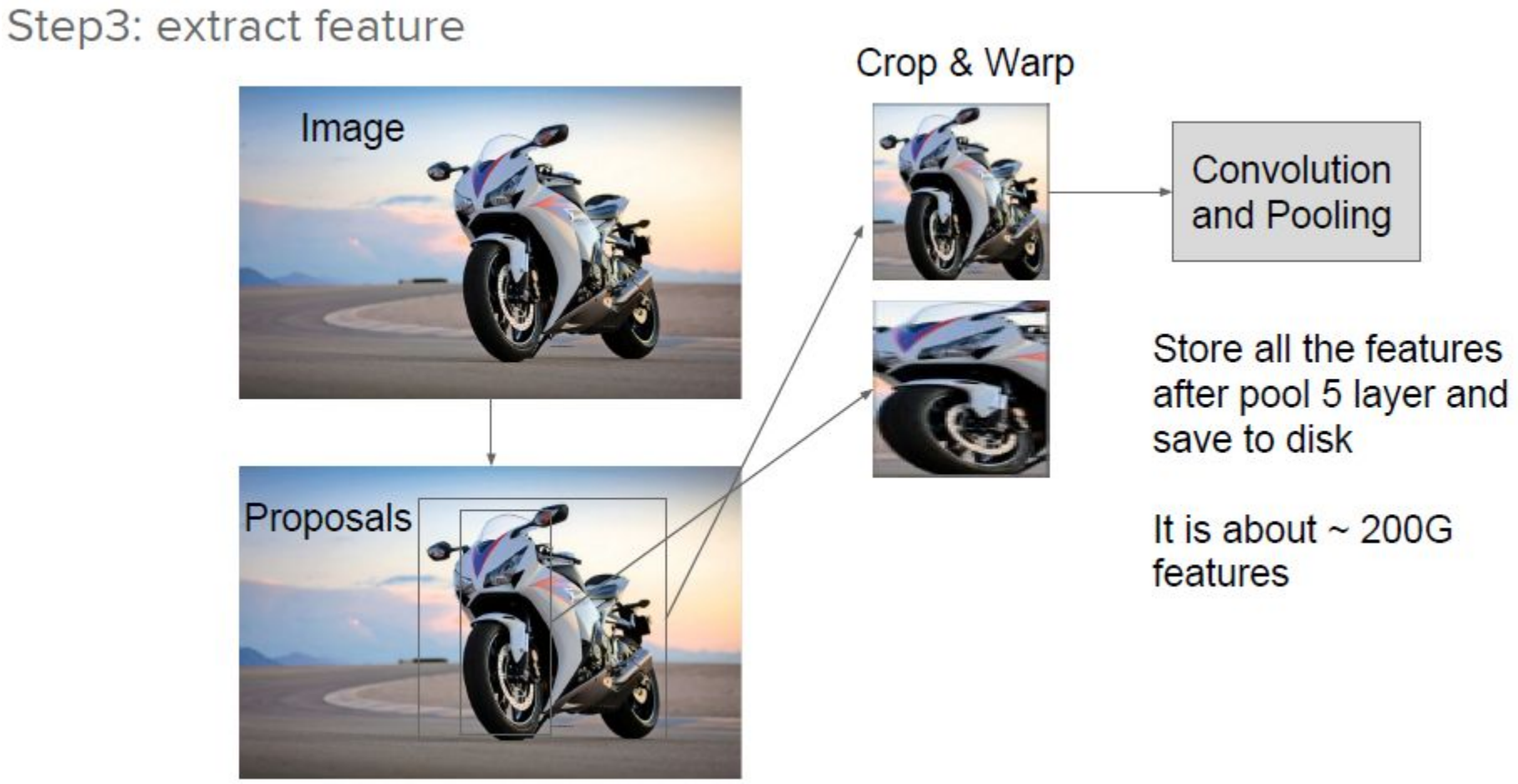
Training RCNN

Step2: focus on 20 classes + 1 background. Remove the last FC layer and replace it with a smaller layer and fine-tune the model using PASCAL VOC dataset



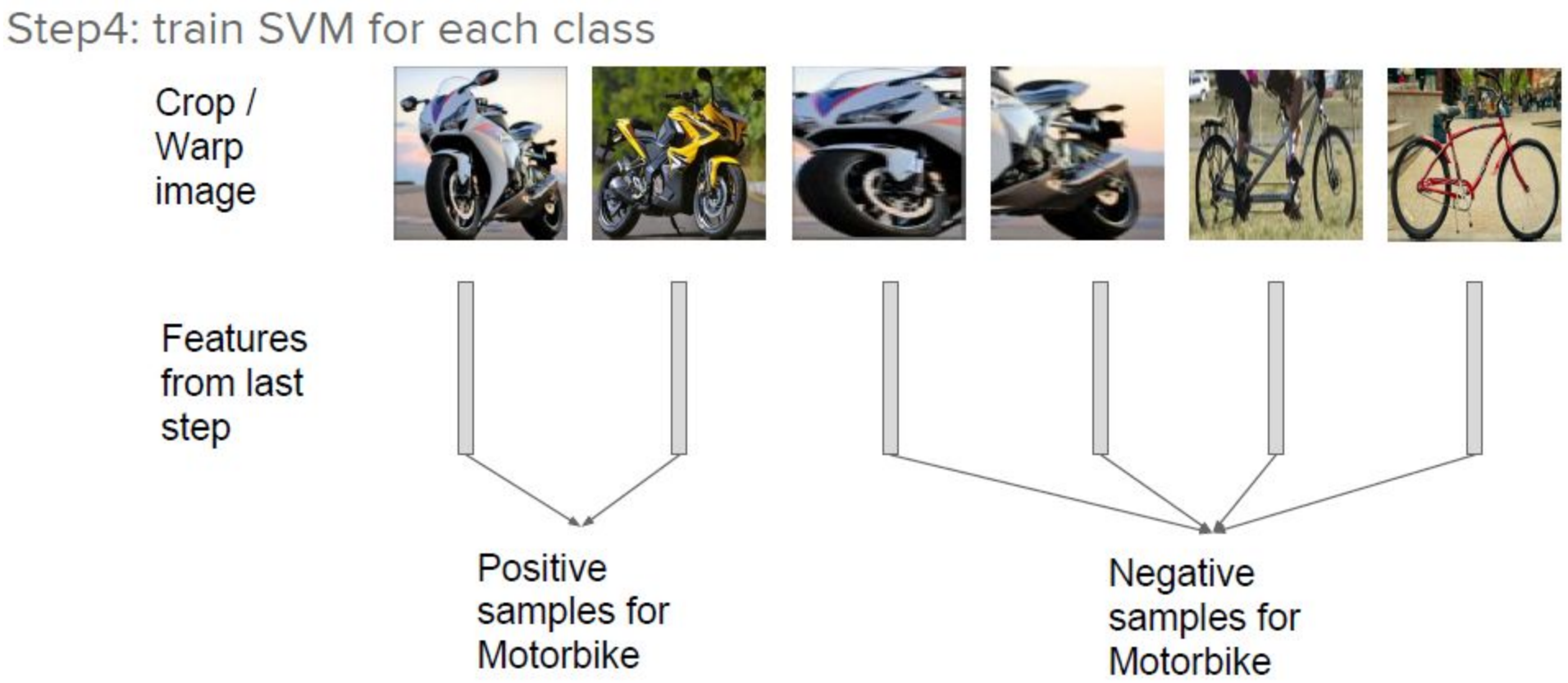
Slide Credit: Sihao Liang et al.

Training RCNN



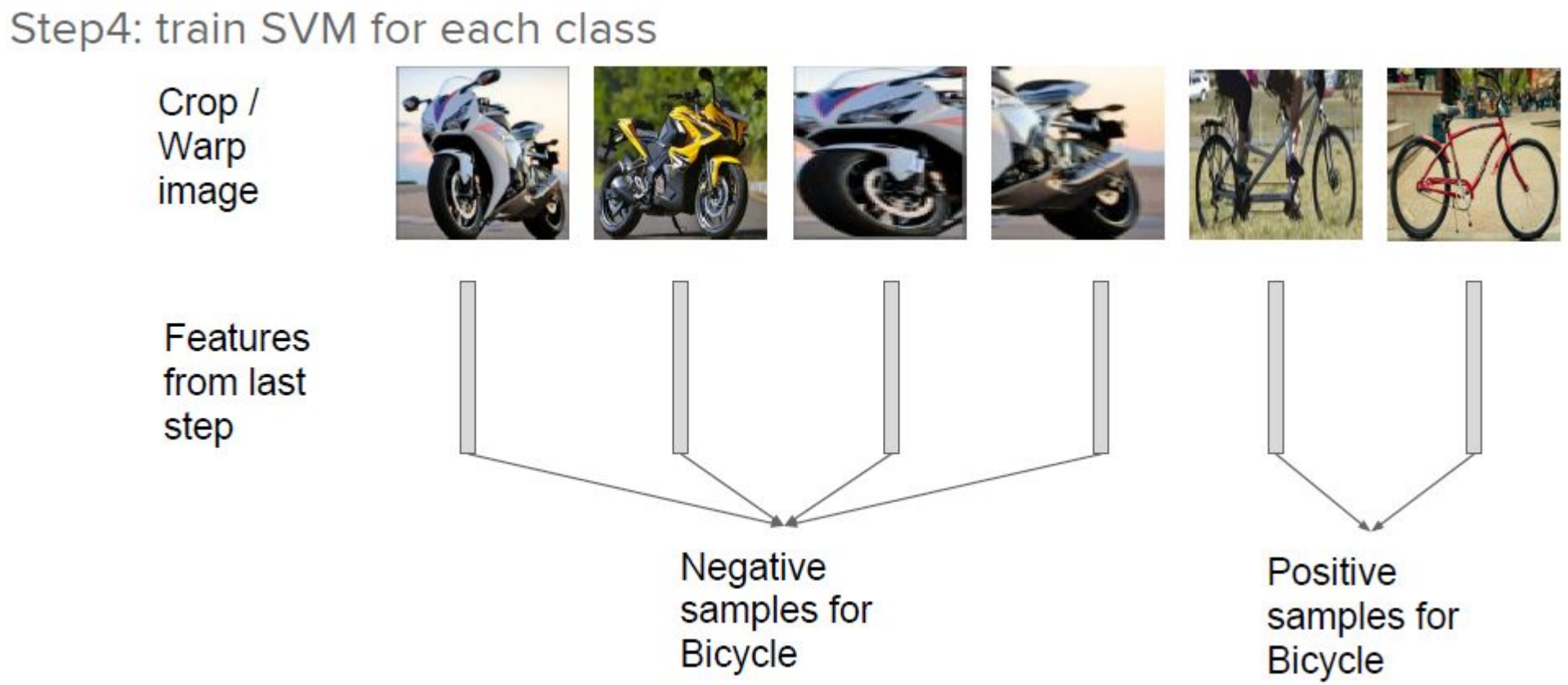
Slide Credit: Sihao Liang et al.

Training RCNN



Slide Credit: Sihao Liang et al.

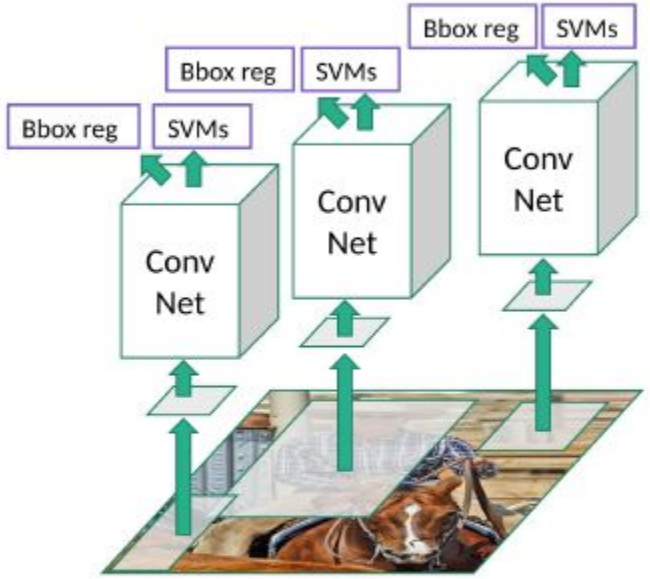
Training RCNN



Slide Credit: Sihao Liang et al.

RCNN: Problems

- Ad hoc training objectives
  - Fine-tune network with softmax classifier (log loss)
  - Train post-hoc linear SVMs (hinge loss)
  - Train post-hoc bounding-box regressions (least squares)
- Training is slow (84h), takes a lot of disk space
- Inference (detection) is slow
  - 47s / image with VGG16 [Simonyan & Zisserman. ICLR15]
  - Fixed by SPP-net [He et al. ECCV14]



Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.
Slide copyright Ross Girshick, 2015; source. Reproduced with permission.

Slide Credit: Fei Fei Li et al. Stanford

Thank you