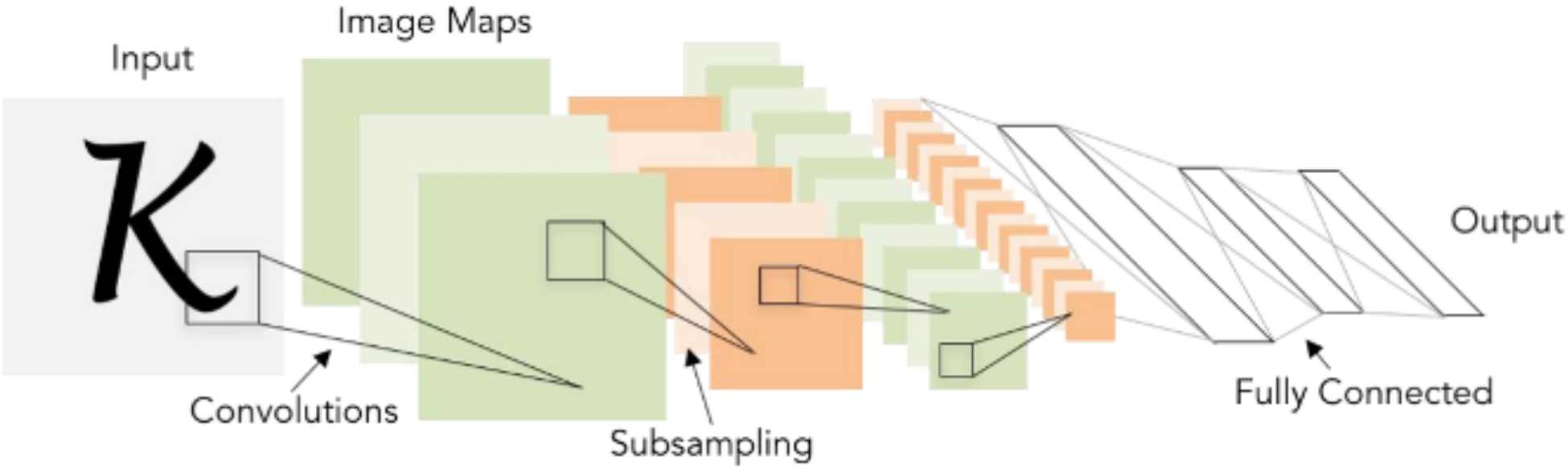
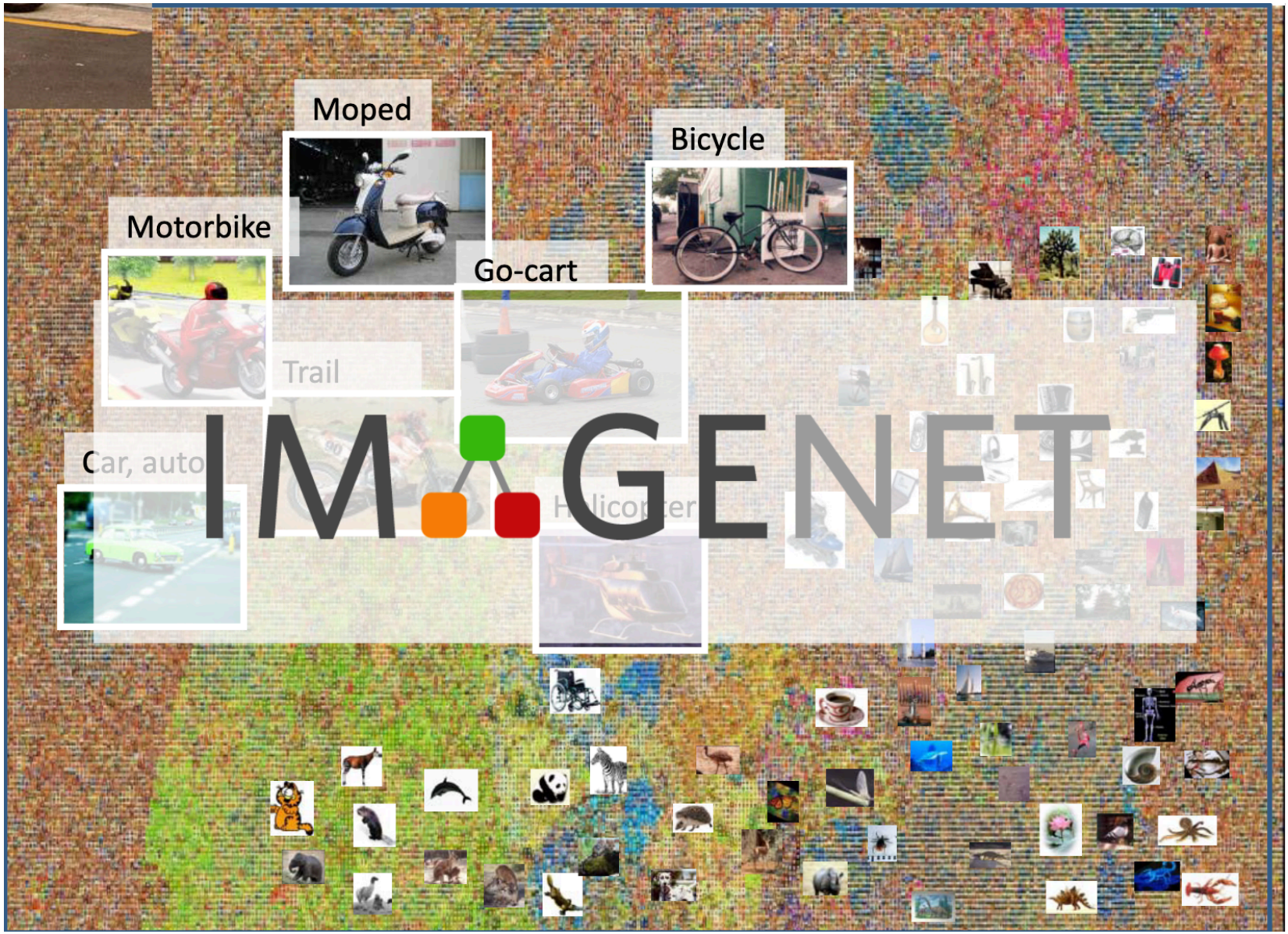
Lecture 17 - Architectures for Recognition

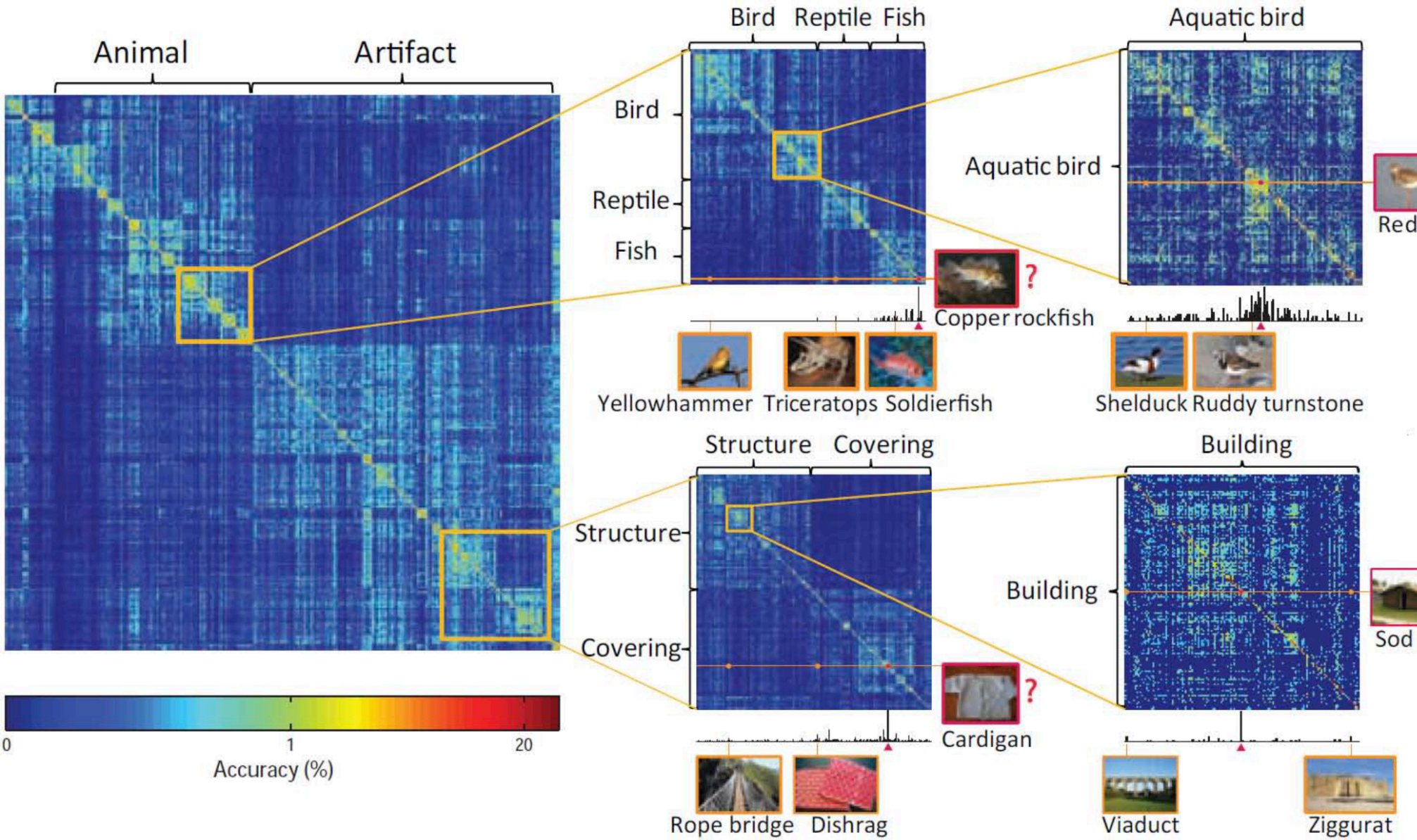
(and other tasks)

Review: LeNet-5

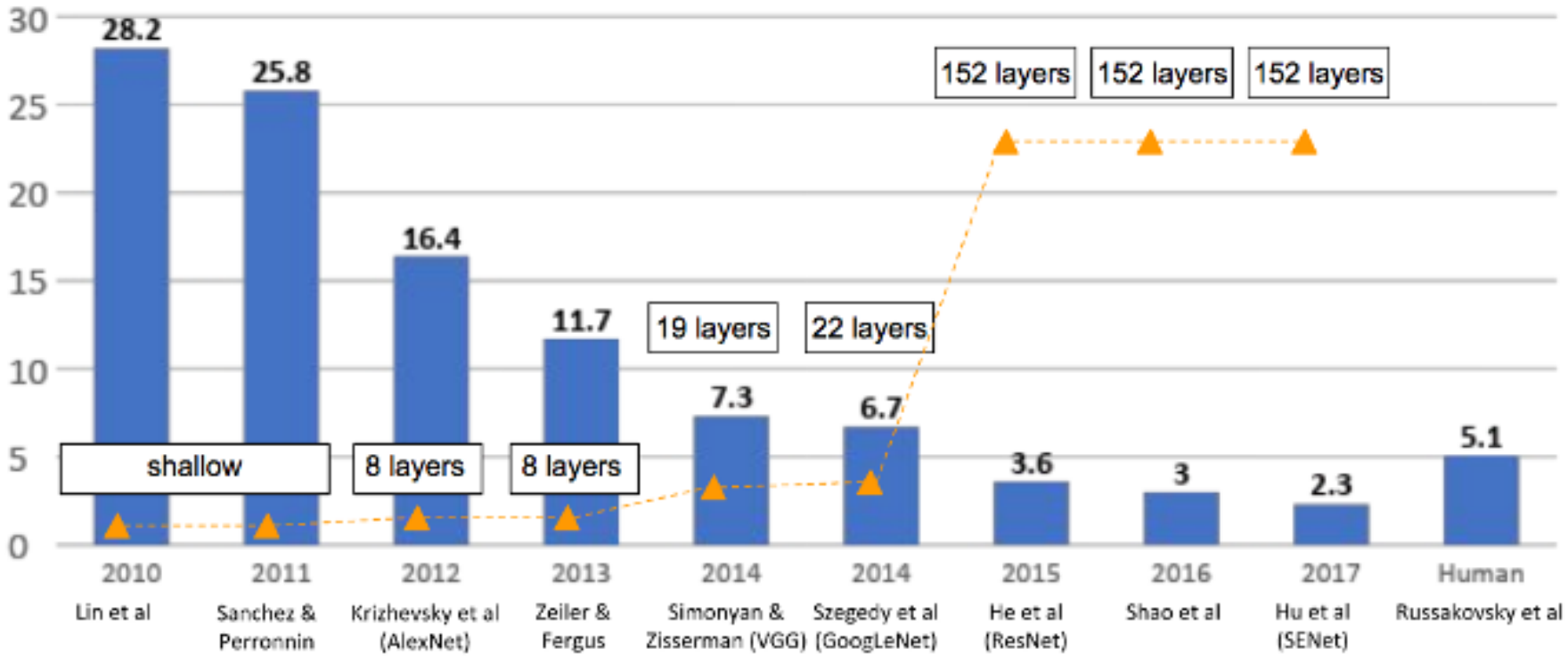
[LeCun et al., 1998]



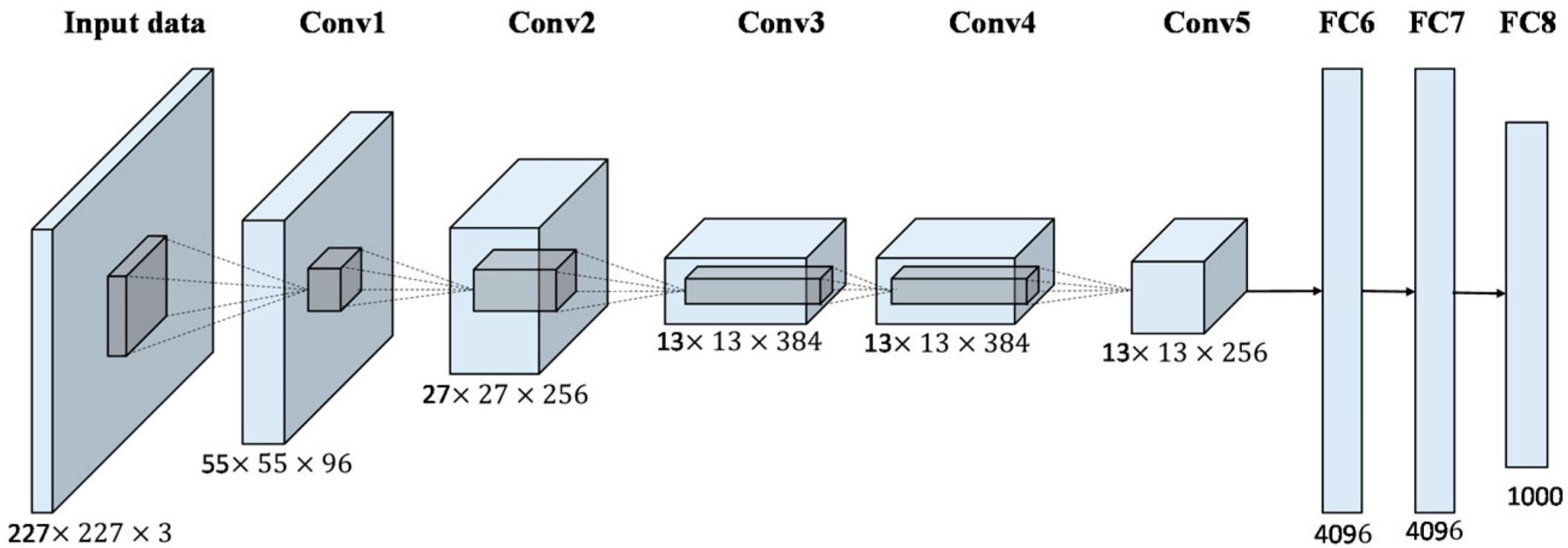

Moped
Bicycle
Motorbike
Go-cart
Trail
IMAGENET


Animal
Artifact
0
1
20
Accuracy (%)
Bird
Reptile
Fish
Yellowhammer
Triceratops
Soldierfish
Copper rockfish
?
Aquatic bird
Shelduck
Ruddy turnstone
Red
Structure
Covering
Rope bridge
Dishrag
Cardigan
Building
Viaduct
Ziggurat
Sod

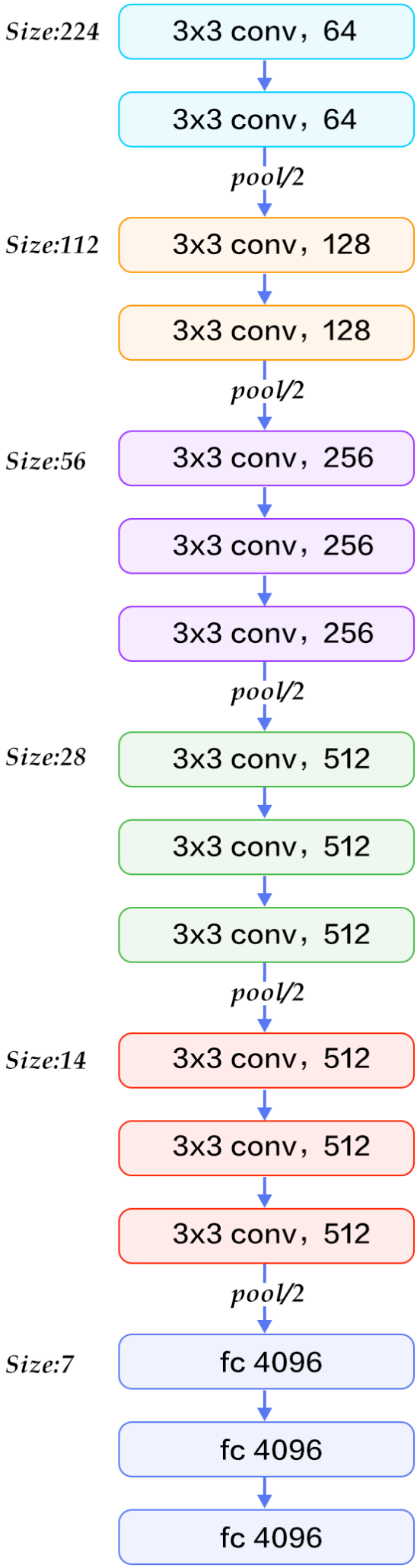
ImageNet Large Scale Visual Recognition Challenge (ILSVRC) winners



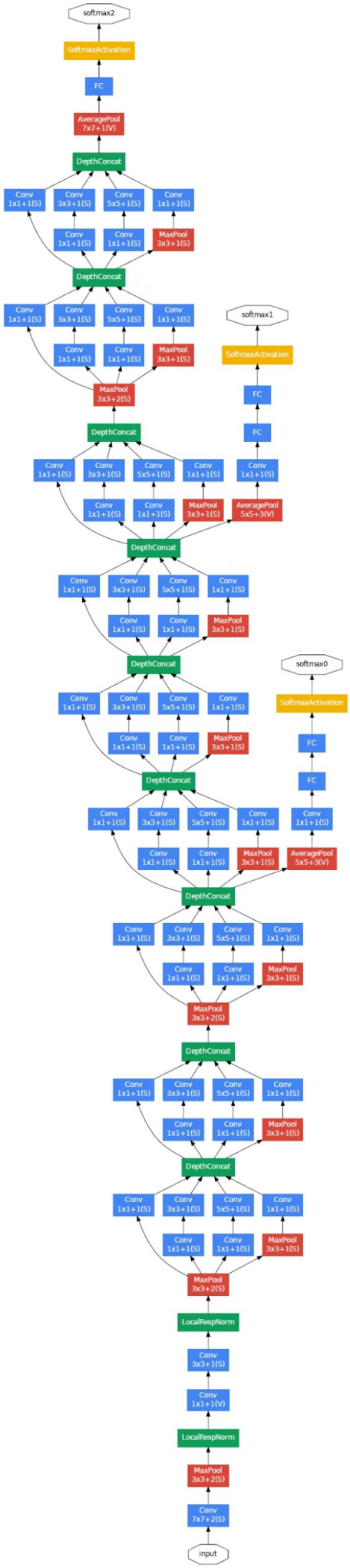

WE NEED TO GO
DEEPER


Input data
Conv1
Conv2
Conv3
Conv4
Conv5
FC6
FC7
FC8
227× 227 × 3
55× 55 × 96
27× 27 × 256
13× 13 × 384
13× 13 × 384
13× 13 × 256
4096
4096
1000


WE NEED TO GO
DEEPER

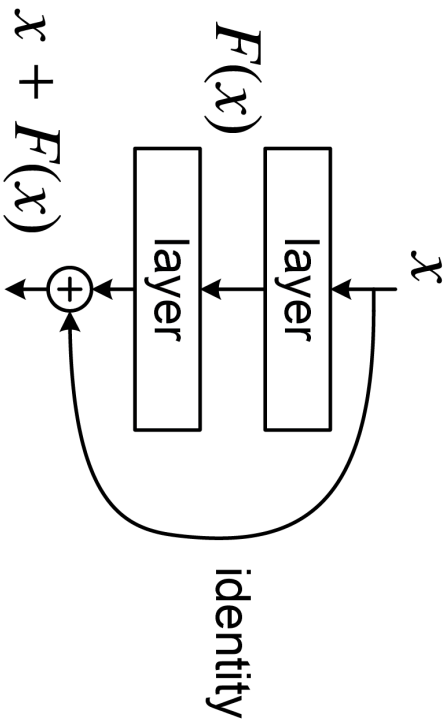

Size:224
3x3 conv，64
3x3 conv，64
pool/2
Size:112
3x3 conv，128
3x3 conv，128
pool/2
Size:56
3x3 conv，256
3x3 conv，256
3x3 conv，256
pool/2
Size:28
3x3 conv，512
3x3 conv，512
3x3 conv，512
pool/2
Size:14
3x3 conv，512
3x3 conv，512
3x3 conv，512
pool/2
Size:7
fc 4096
fc 4096
fc 4096

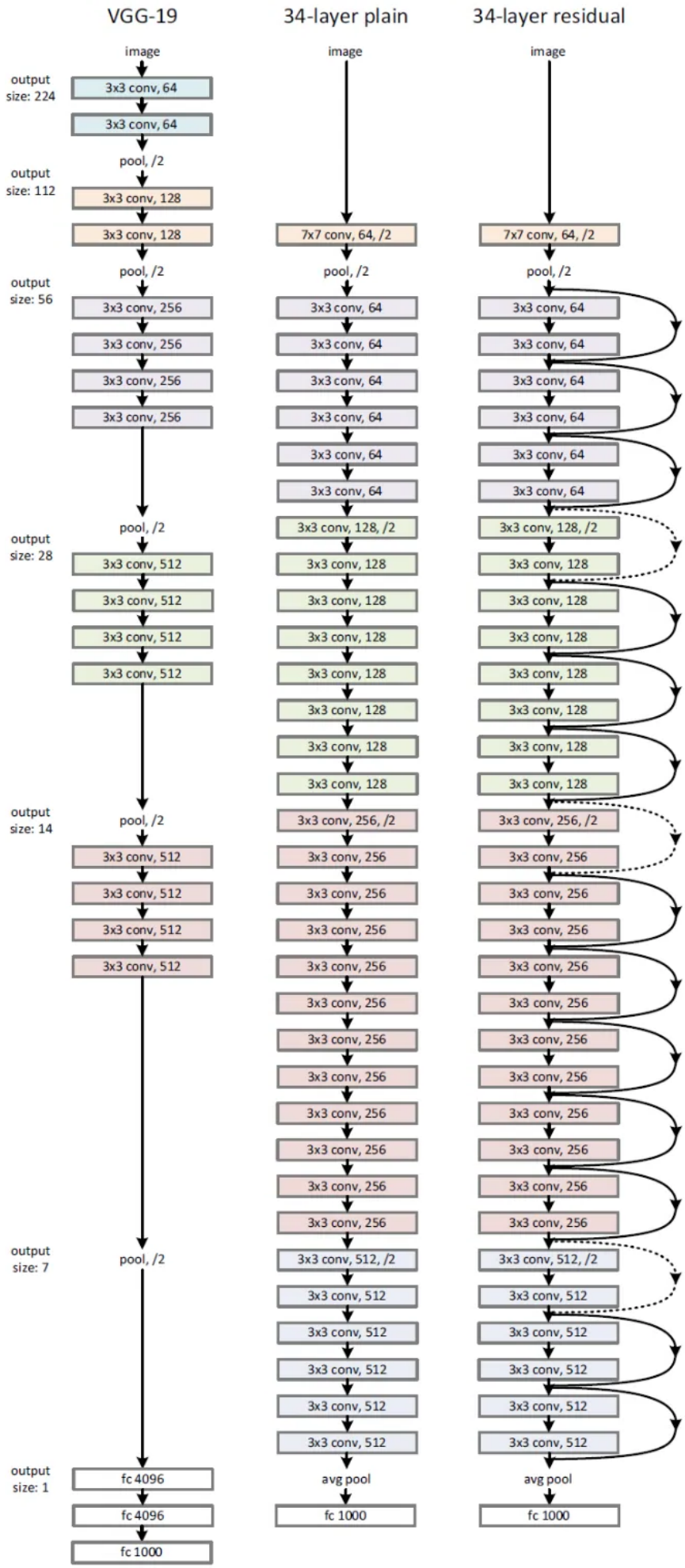

WE NEED TO GO
DEEPER


softmax2
SoftmaxActivation
FC
AveragePool 7x7+1(V)
DepthConcat
Conv 1x1+1(S)
Conv 3x3+1(S)
Conv 5x5+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
MaxPool 3x3+1(S)
DepthConcat
Conv 1x1+1(S)
Conv 3x3+1(S)
Conv 5x5+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
MaxPool 3x3+1(S)
MaxPool 3x3+2(S)
softmax1
SoftmaxActivation
FC
FC
DepthConcat
Conv 1x1+1(S)
Conv 3x3+1(S)
Conv 5x5+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
MaxPool 3x3+1(S)
AveragePool 5x5+3(V)
DepthConcat
Conv 1x1+1(S)
Conv 3x3+1(S)
Conv 5x5+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
MaxPool 3x3+1(S)
DepthConcat
softmax0
SoftmaxActivation
FC
FC
Conv 1x1+1(S)
Conv 3x3+1(S)
Conv 5x5+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
MaxPool 3x3+1(S)
AveragePool 5x5+3(V)
DepthConcat
Conv 1x1+1(S)
Conv 3x3+1(S)
Conv 5x5+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
MaxPool 3x3+1(S)
MaxPool 3x3+2(S)
DepthConcat
Conv 1x1+1(S)
Conv 3x3+1(S)
Conv 5x5+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
MaxPool 3x3+1(S)
DepthConcat
Conv 1x1+1(S)
Conv 3x3+1(S)
Conv 5x5+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
MaxPool 3x3+1(S)
MaxPool 3x3+2(S)
LocalRespNorm
Conv 3x3+1(S)
Conv 1x1+1(V)
LocalRespNorm
MaxPool 3x3+2(S)
Conv 7x7+2(S)
input


AFP


WE NEED TO GO
DEEPER

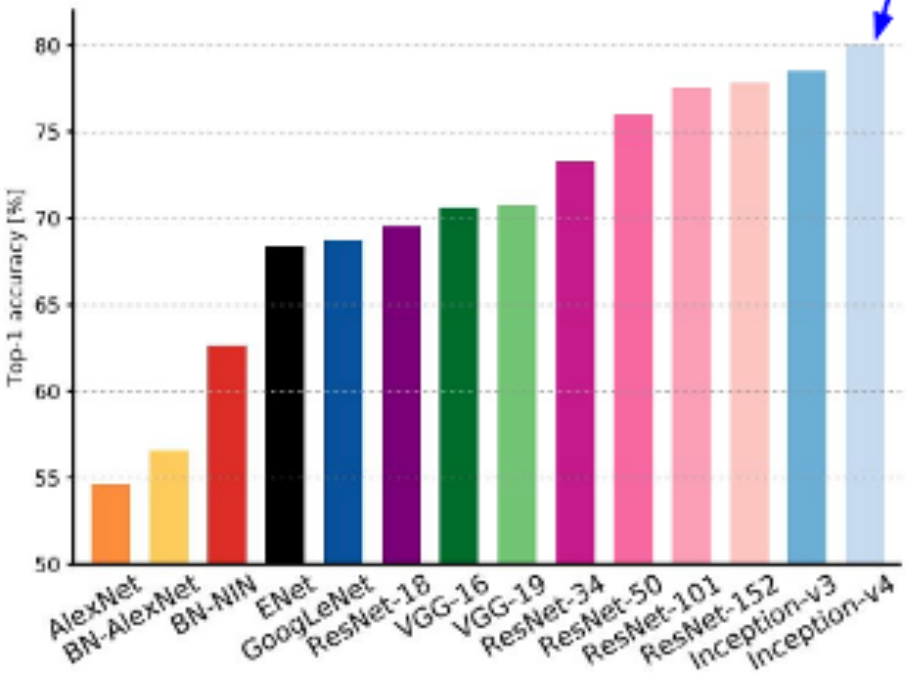

$x + F(x)$
$F(x)$
layer
layer
x
identity

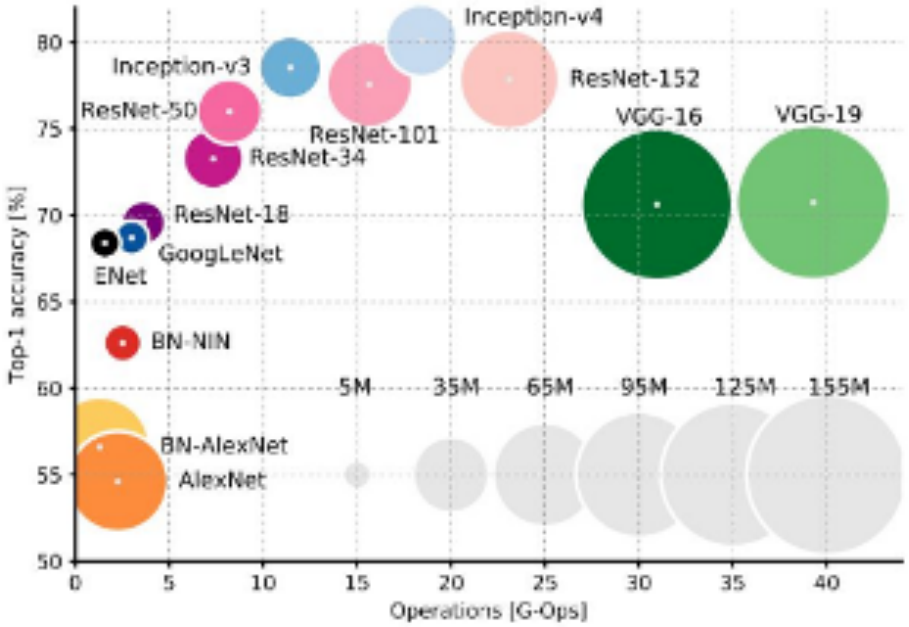

VGG-19
34-layer plain
34-layer residual
image
output size: 224
3x3 conv, 64
3x3 conv, 64
pool, /2
output size: 112
3x3 conv, 128
3x3 conv, 128
pool, /2
output size: 56
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
pool, /2
output size: 28
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
pool, /2
output size: 14
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
pool, /2
output size: 7
fc 4096
fc 4096
fc 1000
7x7 conv, 64, /2
pool, /2
3x3 conv, 64
3x3 conv, 128, /2
3x3 conv, 128
3x3 conv, 256, /2
3x3 conv, 256
3x3 conv, 512, /2
3x3 conv, 512
avg pool
fc 1000

Do we?

Comparing complexity…

Inception-v4: Resnet + Inception!



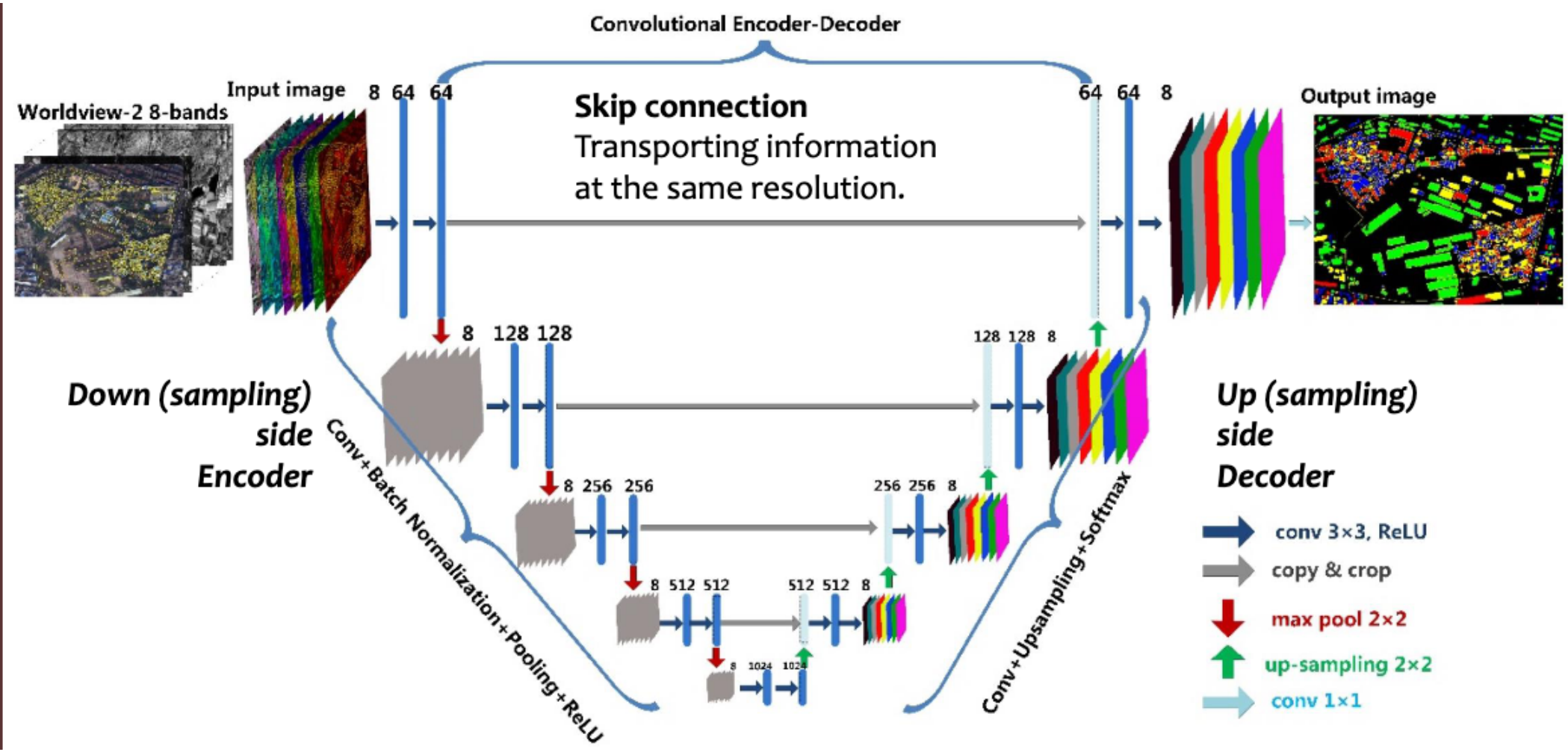


An Analysis of Deep Neural Network Models for Practical Applications, 2017.

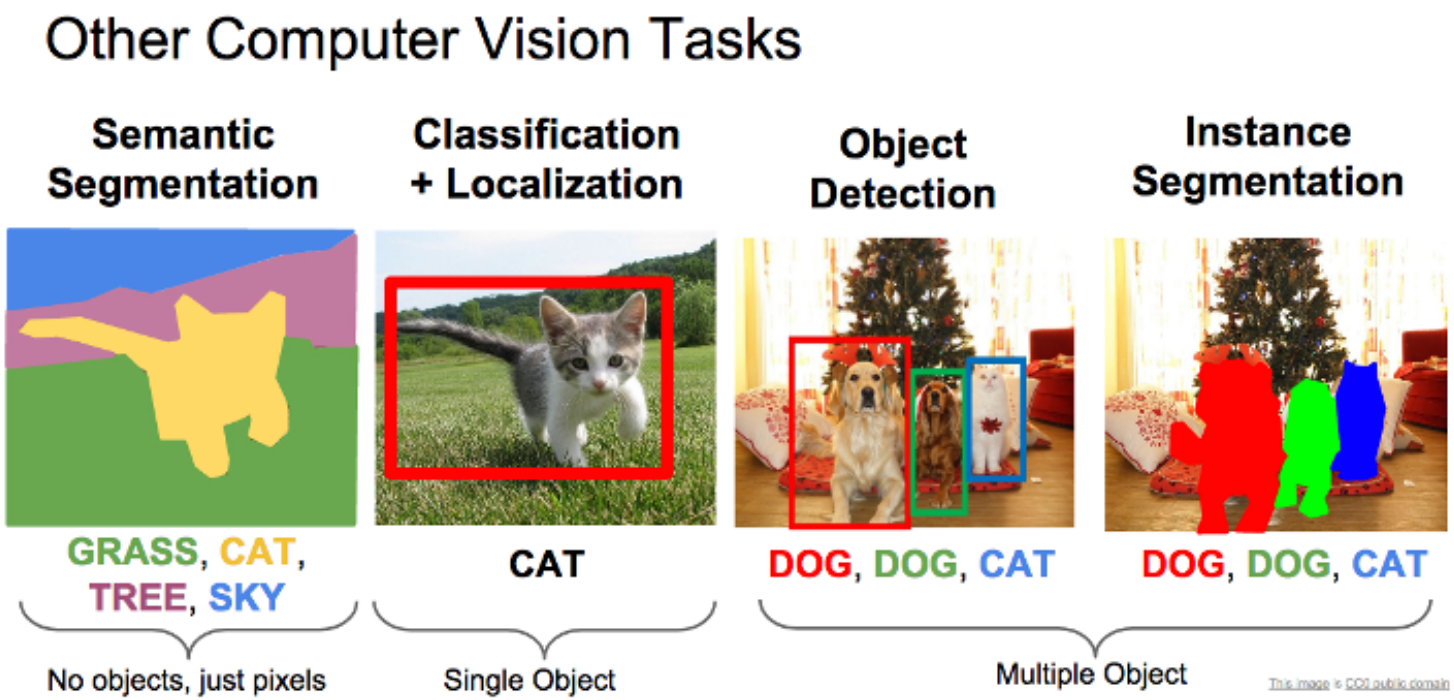
Okay but the data...

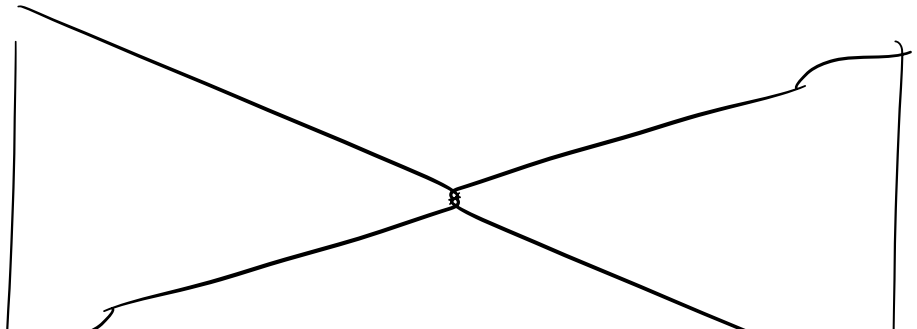
Transfer Learning / finetuning

UNet - Architecture for Dense Prediction Tasks



What about not image recognition?





Other vision tasks

- Optical Flow
  https://flowseek25.github.io/
- Feature Matching, Stereo Matching
  https://stereoanywhere.github.io/
- 3D Reconstruction
  https://github.com/naver/dust3r
  https://github.com/naver/mast3r
  https://vgg-t.github.io/
- Image and Video Segmentation
  https://ai.meta.com/sam3/
-

Architecture of the day: Transformer