Convolutional Neural Networks on Image Recognition

Elliott Huangfu

Oct.4 2018

Image recognition - Applications

Google / iCloud Photos

- Recognize objects and faces
- Semantic search for images

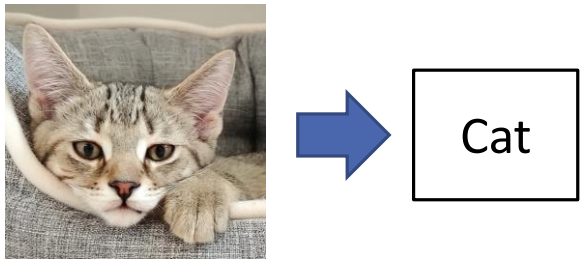
Content filtering

- Facebook scans 2 billion photos every day
- Looking for
  - Objectionable content, violence, pornography
  - Generate label and captions, for visually impaired
  - Generate tags, for interest recommender system
  - Face recognition

Tasks: Classification & Detection

Classification

- “Contextual Image Classification”



Object Detection

- Classification + localization

Table of Content

The Fundamentals

Convolution, Pooling

Problem with multi-layer perceptron

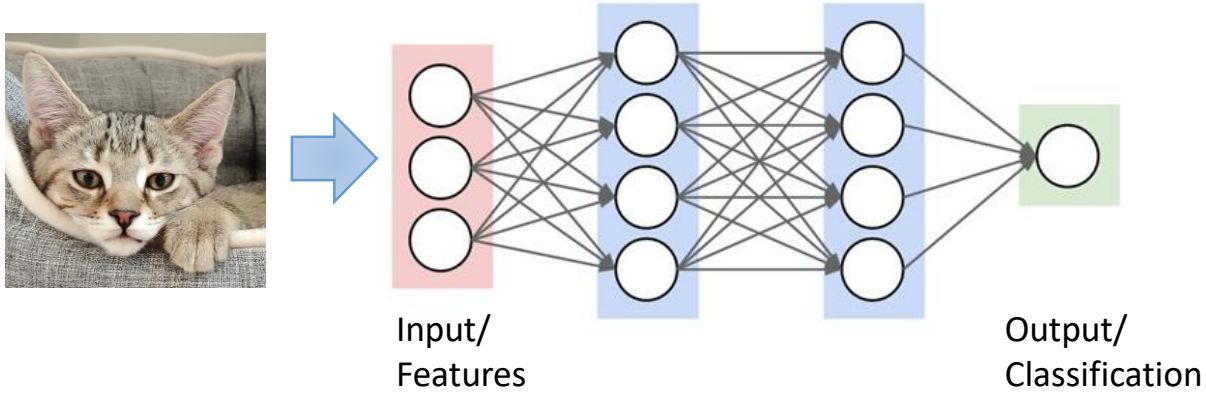
Huge input features

- 64x64x3 = 12,288
- 480x480x3 = 691,200

Difficult to train

Overfitting

- Sensitive to pattern size and position



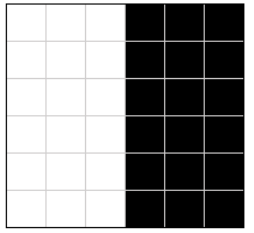
Convolution

- Act as feature detectors (edges, curves)
- Reduce the parameters between layers (parameter sharing)

Matrix Convolution

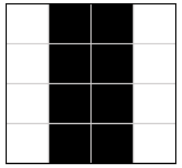
Convolve Operation

- “Filter”
- Many settings
- Feature extraction



6x6

Vertical edge detector



4x4

0	0	0	9	9	9
0	0	0	9	9	9
0	0	0	9	9	9
0	0	0	9	9	9
0	0	0	9	9	9
0	0	0	9	9	9

*

-1	0	1
-1	0	1
-1	0	1

Vertical edge detector

=

0	27	27	0
0	27	27	0
0	27	27	0
0	27	27	0

https://en.wikipedia.org/wiki/Kernel_(image_processing)
https://docs.gimp.org/en/plug-in-convmatrix.html

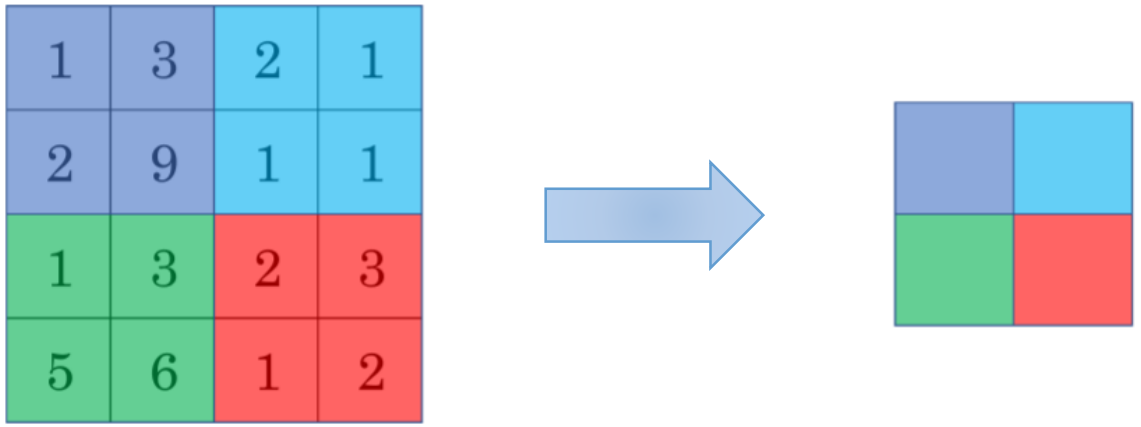
Pooling

Down-sampling

- No Parameters

Max pooling

Average pooling



Evolution of Algorithms

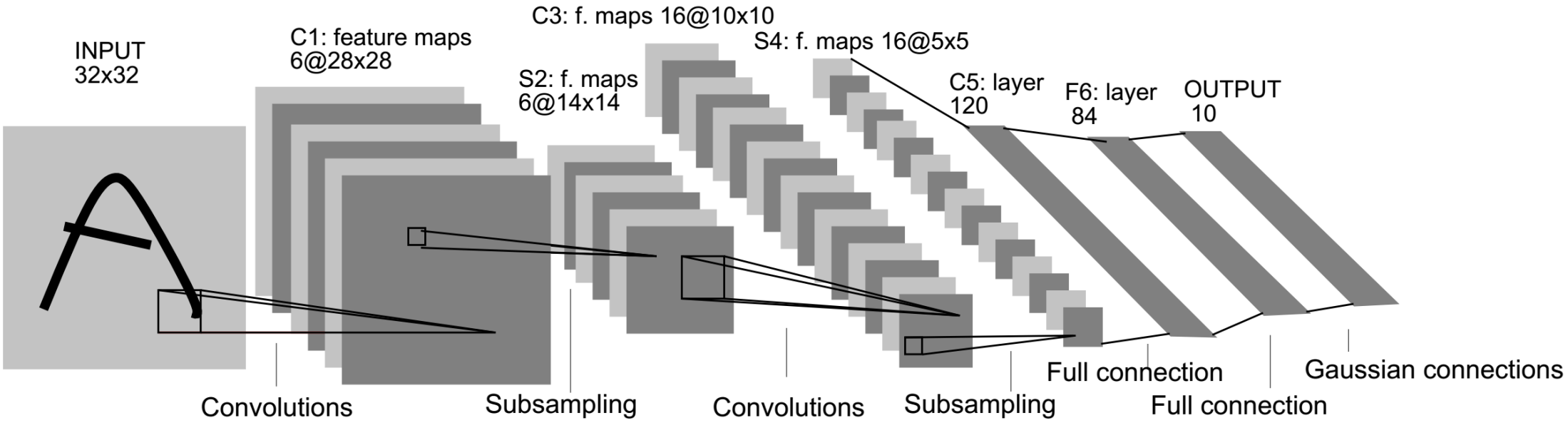
LeNet-5, AlexNet, VGG, GoogLeNet, ResNet, …

LeNet-5 – the Pioneer (1998)

32x32 image, grayscale

2 Conv + 3 FC layers, ~60k parameters.

Hand written digits recognition



Yann Lecun, et al., “Gradient-Based Learning Applied to Document Recognition.” In Proceedings of the IEEE, 2278–2324, 1998.

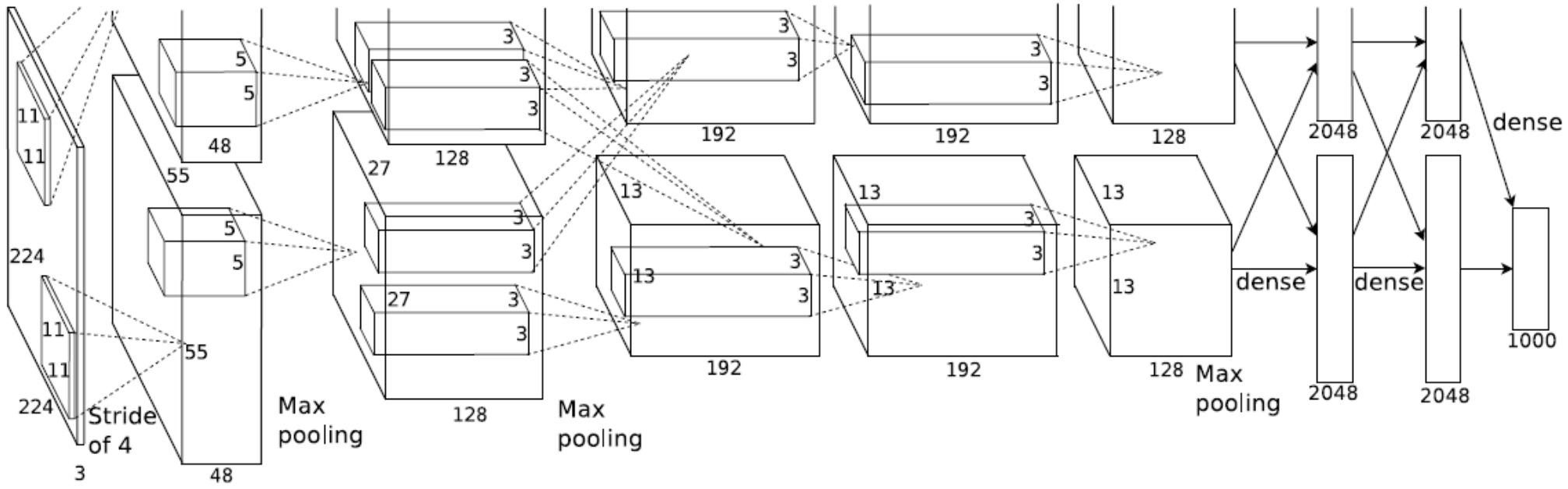
AlexNet (2012)

224x224x3 image, RGB

5 conv + 3 FC layers, ~60m parameters

ReLU / data augmentation / dropout

Error Rate on ImageNet 2011	
AlexNet	15.3%
2nd Best Algm.	26.2%



Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “ImageNet Classification with Deep Convolutional Neural Networks.” In Advances in Neural Information Processing Systems 25, 1097–1105, 2012.

VGG-16 / 19 (2014) - Deeper

Unified Framework

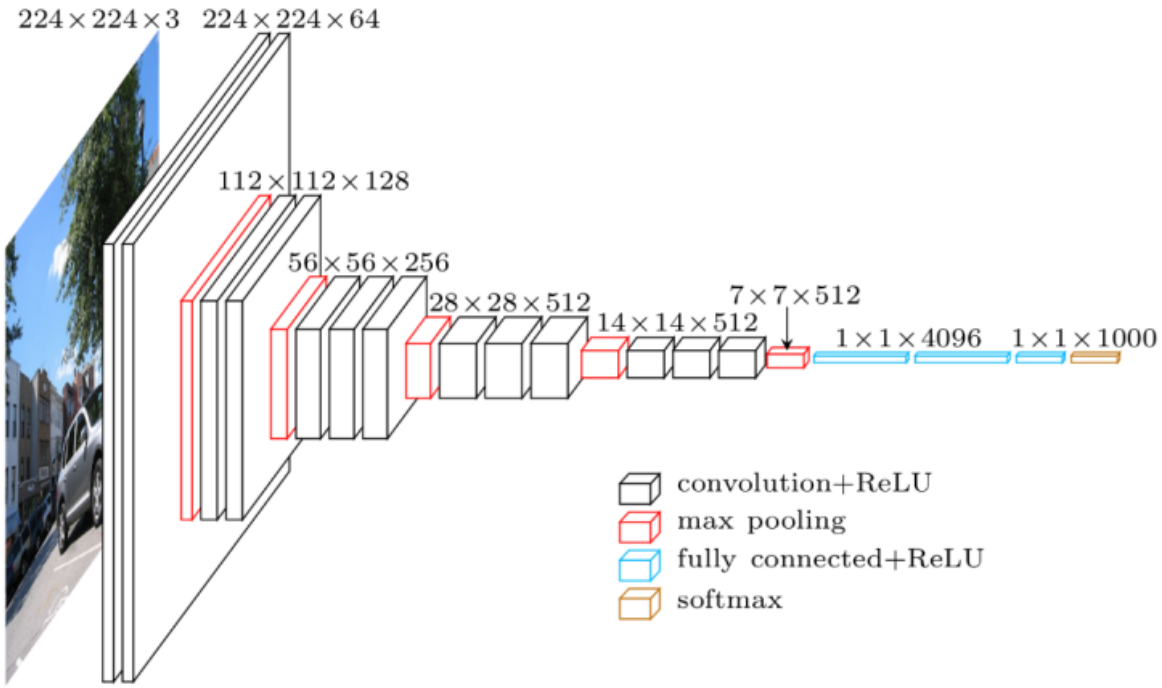
6 variations, 130~140m parameters

- VGG-16: 13 Conv layers, 3 FC layers
- VGG-19: 16 Conv layers, 3 FC layers

"**Depth is beneficial**"

Performance increment slows down as network goes deeper

- Gradient vanish



Karen Simonyan, and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." ArXiv:1409.1556 [Cs], 2014.

GoogLeNet (2014) – Wilder

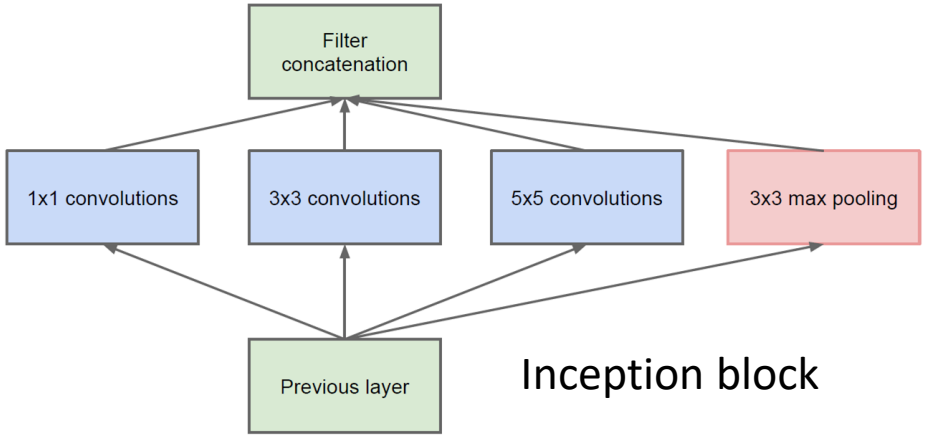
"Inception block"

- Let the algorithm choose its own filters.

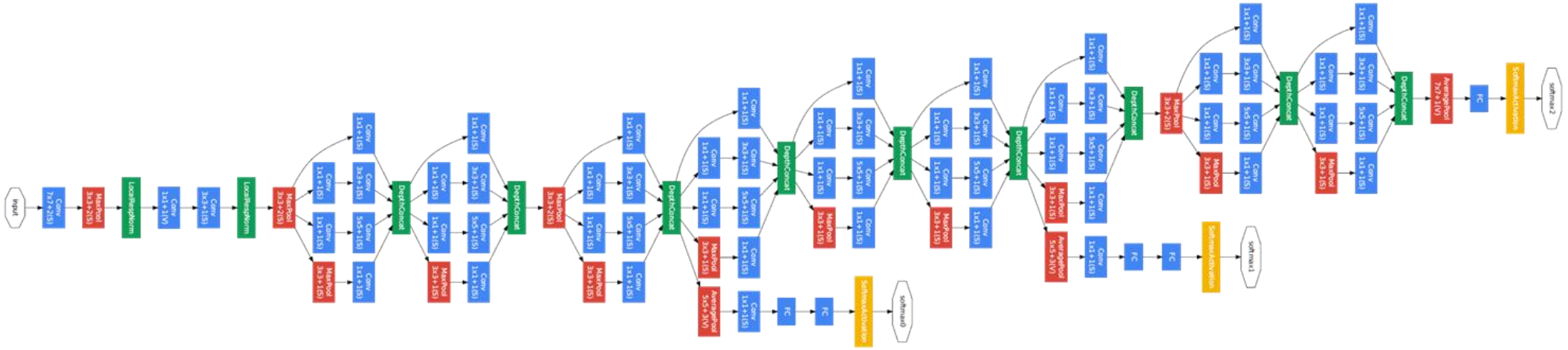
21 effective conv layers + 1 FC layer

6.8m parameters

Sparser, but wider network yields better result.



Inception block



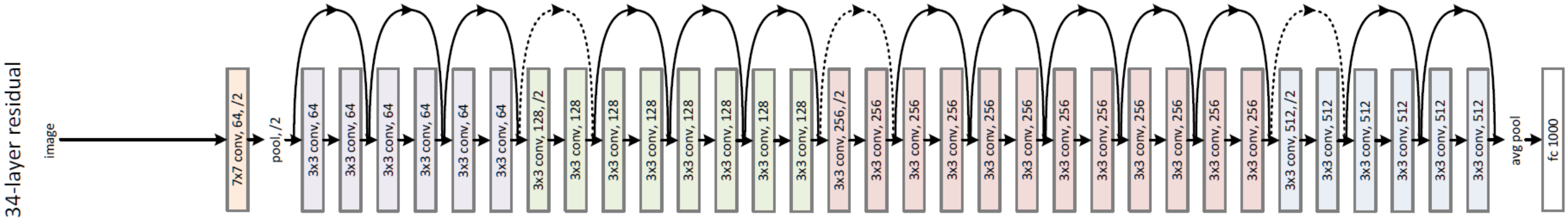
Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going Deeper with Convolutions." ArXiv:1409.4842 [Cs], September 16, 2014

ResNet - Deep Residual Network (2015)

Solved gradient vanish problem

Formatted structure, scalable, 50/101/152/1202 layers

1.7m parameters for 110-layer



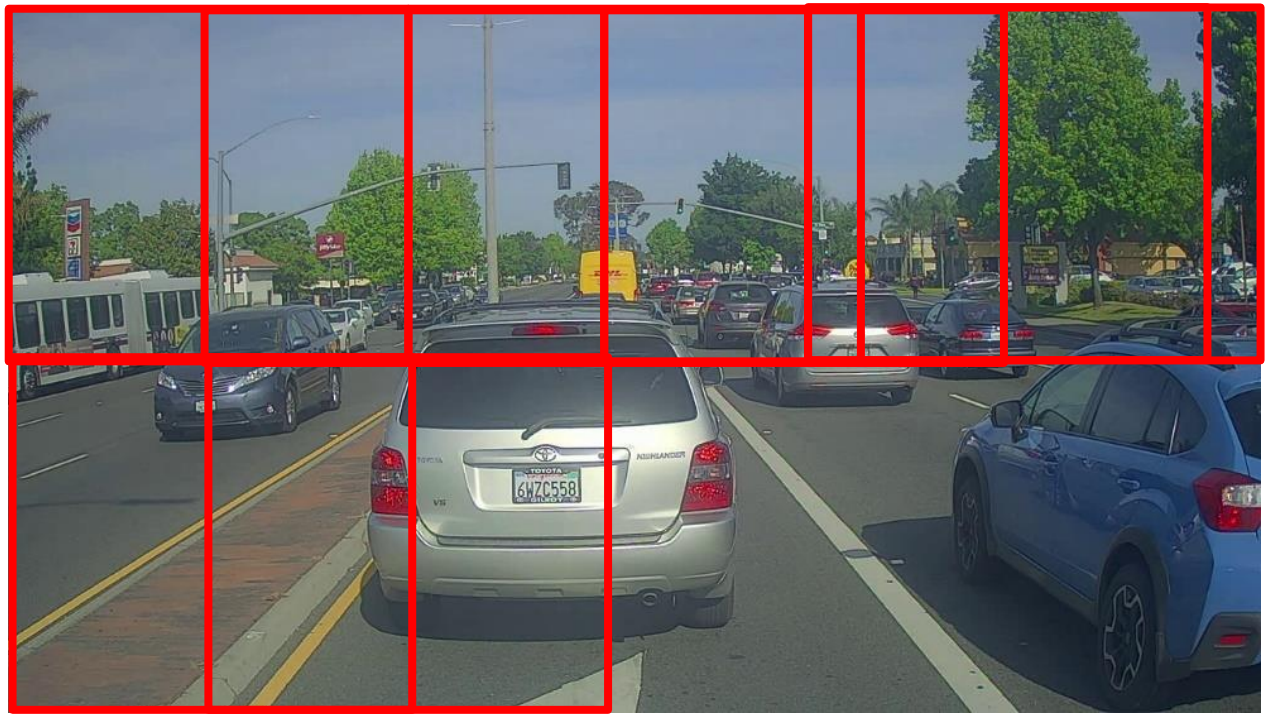
Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep Residual Learning for Image Recognition.” ArXiv:1512.03385 [Cs], December 10, 2015.

Detection

Sliding window

- Shape, size, step
- Compute-intensive

Regional CNN (R-CNN)



YOLO – Efficient Detection

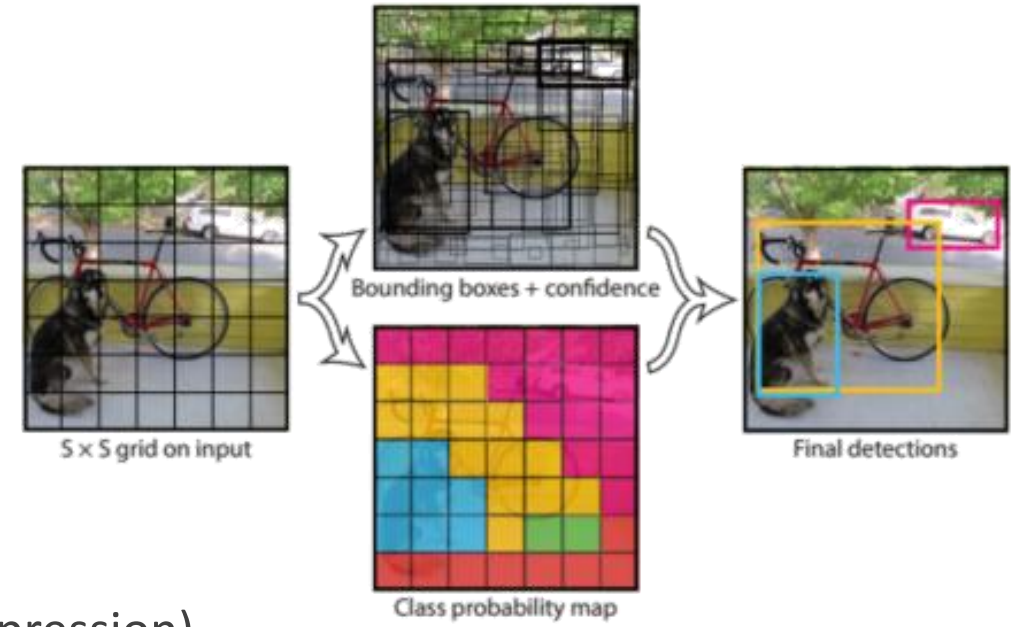
"You Only Look Once"

- End-to-end detection

Classification model (v2): similar to VGG-19

Direct location prediction

- S x S grid cells
- Each cell predicts object bonding boxes
- Combine results and remove invalids (non-max suppression)



Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi. "You Only Look Once: Unified, Real-Time Object Detection." ArXiv:1506.02640 [Cs], June 8, 2015.
Redmon, Joseph, and Ali Farhadi. "YOLO9000: Better, Faster, Stronger." ArXiv:1612.08242 [Cs], December 25, 2016.

YOLO – Efficient Detection

"You Only Look Once"

Classification model (v2): similar to VGG-19

Direct location prediction
- S x S grid cells
- Each cell predicts object bonding boxes
- Combine results and remove duplicates (non-max suppress

YOLO v2 speed on a Titan X GPU
- 544x544 input: **40 FPS**
- 288x288 input: **90 FPS**

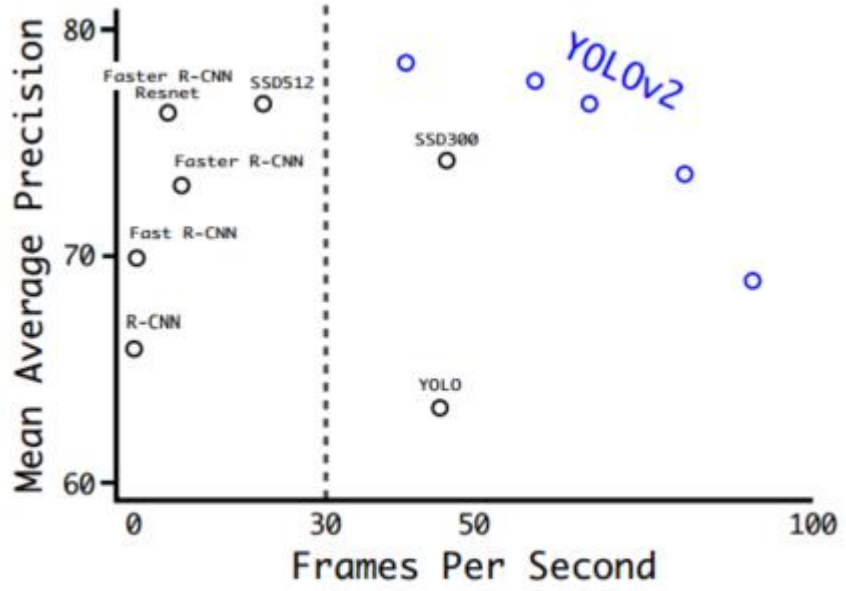


Figure 4: Accuracy and speed on VOC 2007.

Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi. "You Only Look Once: Unified, Real-Time Object Detection." ArXiv:1506.02640 [Cs], June 8, 2015.
Redmon, Joseph, and Ali Farhadi. "YOLO9000: Better, Faster, Stronger." ArXiv:1612.08242 [Cs], December 25, 2016.

Evaluation

Precision, Recall, ILSVRC

Performance Metrics

Classification

- Top-1 / Top-5 **Error Rate** $= \frac{prediction\ errors}{N\ of\ images}$

Detection

- Accurate: how to define?

Label:
Sports car

Prediction:
Car wheel
Sports car
Race car
Convertible
Washer

Example of ImageNet images: https://cs.stanford.edu/people/karpathy/ilsvrc/

Performance Metrics

Precision & Recall (simplest form)

Precision

$$P = \frac{TP}{TP + FP}$$

Recall

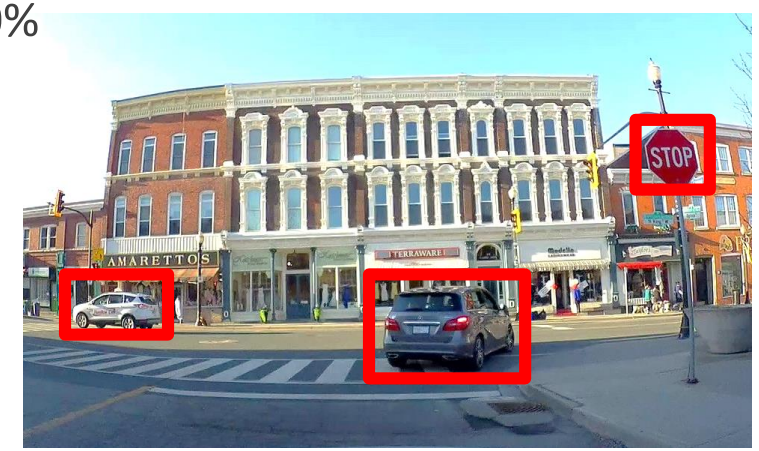
$$R = \frac{TP}{TP + FN}$$

		Ground Truth of Test Data	
		1	**0**
Prediction	**1**	TP True Positive	FP False Positive
	0	FN False Negative	TN True Negative

- E.g. Ground truth 50:50, if prediction all 1, then P=50%, R=100%
- Ground truth all 1, if prediction 50:50, then P=100%, R=50%

Combined metrics

- F1 score = 2*P*R/(P+R)
- Precision-Recall Curve
- AP: Average Precision over different levels of recall



https://en.wikipedia.org/wiki/Precision_and_recall
Jensen, M. B., K. Nasrollahi, and T. B. Moeslund. “Evaluating State-of-the-Art Object Detector on Challenging Traffic Light Data.” In *2017 IEEE Conference on CVPRW*, 882–88, 2017.

ILSVRC

ImageNet Large Scale Visual Recognition Challenge

- 1000 object classes, 1m+ images

	Classification		Detection	
	Winner (algorithm)	**Error Rate**	**Winner (algorithm)**	**Average Precision**
2010	NEC	28.2%		
2011	XRCE	25.8%		
2012	SuperVision (**AlexNet**)	16.4%		
2013	Clarifai (**AlexNet**)	11.74%	UvA	22.6%
2014	**GoogLeNet**	6.66%	NUS	37.2%
2015	MSRA (**ResNet**)	3.57%	MSRA (**ResNet**)	62.1%
2016	Trimps-Soushen (combined)	2.99%	CUImage (GBD-Net)	66.3%
2017	WMW (SE-ResNet)	2.25%	BDAT (Attention Net)	73.1%

Human-Level Performance

Label:
chest

GoogLeNet Prediction:
honeycomb
French loaf
stole
thimble
velvet

Example of ImageNet images: https://cs.stanford.edu/people/karpathy/ilsvrc/

Summary

Breakthroughs in computer vision have been made with the introduction of CNN

- Surpassed human level performance in certain areas
- Lacks common sense based on historical information

A set of tools is well established for different applications

- AlexNet
- VGG-16/19
- GoogLeNet
- ResNet
- R-CNN
- YOLO

Thank you!