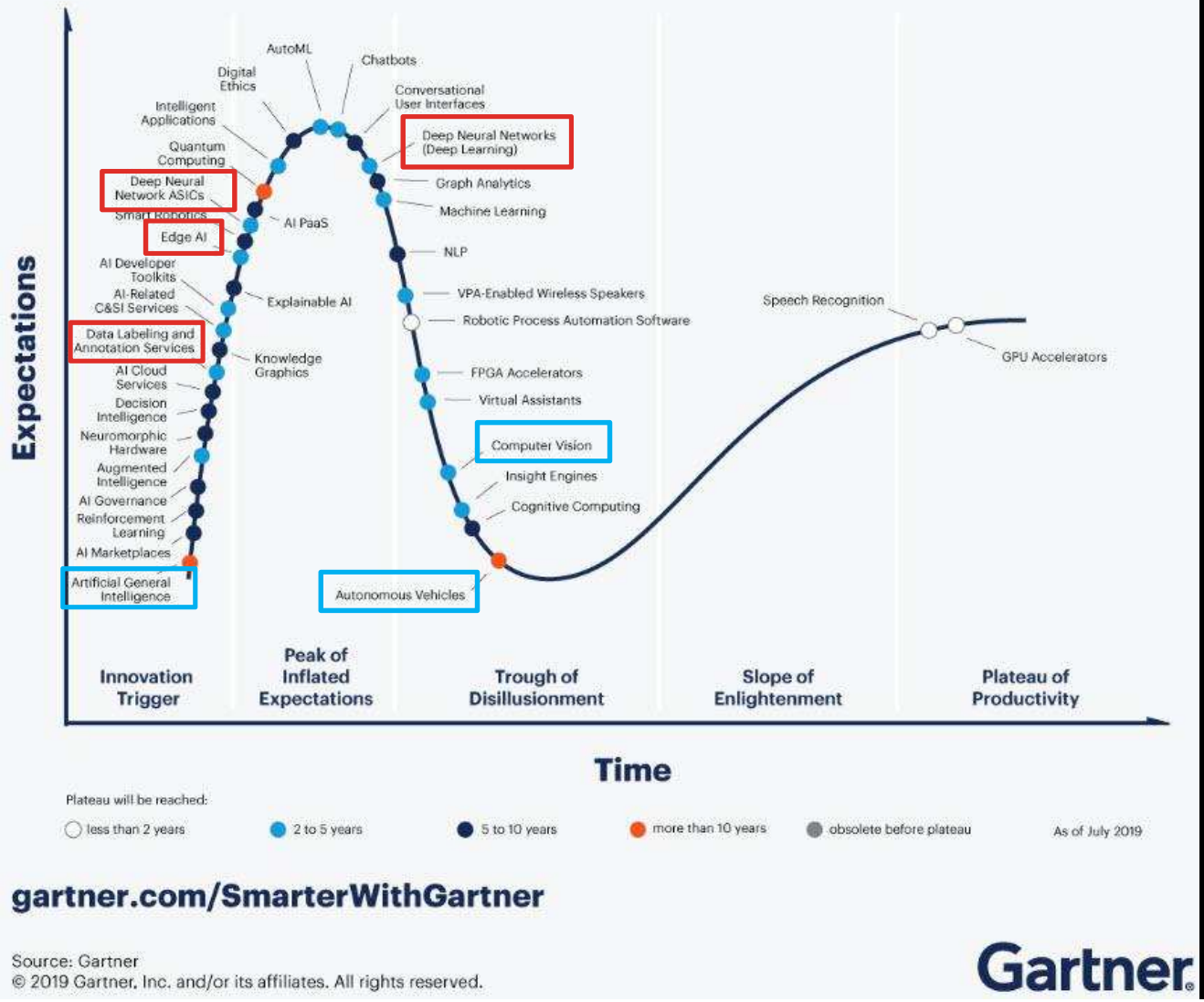
BASICS OF AI VISUAL ALGORITHMS

Kuan-Hung Chen/陳冠宏

Gartner Hype Cycle for Artificial Intelligence, 2019



gartner.com/SmarterWithGartner

Gartner

- Deep Neural Network (Deep Learning)
- Deep Neural Network ASICs
- Edge AI
- Data Labeling and Annotation Services

OUTLINE

- What are AI visual algorithms?
  - How does a computer classify pictures?
  - How does a computer detect objects?
  - What else can AI visual algorithms do?
- How to evaluate performance of AI visual algorithms?

OUTLINE

- What are AI visual algorithms?
  - How does a computer classify pictures?
  - How does a computer detect objects?
  - What else can AI visual algorithms do?
- How to evaluate performance of AI visual algorithms?

HOW DOES A COMPUTER CLASSIFY PICTURES?

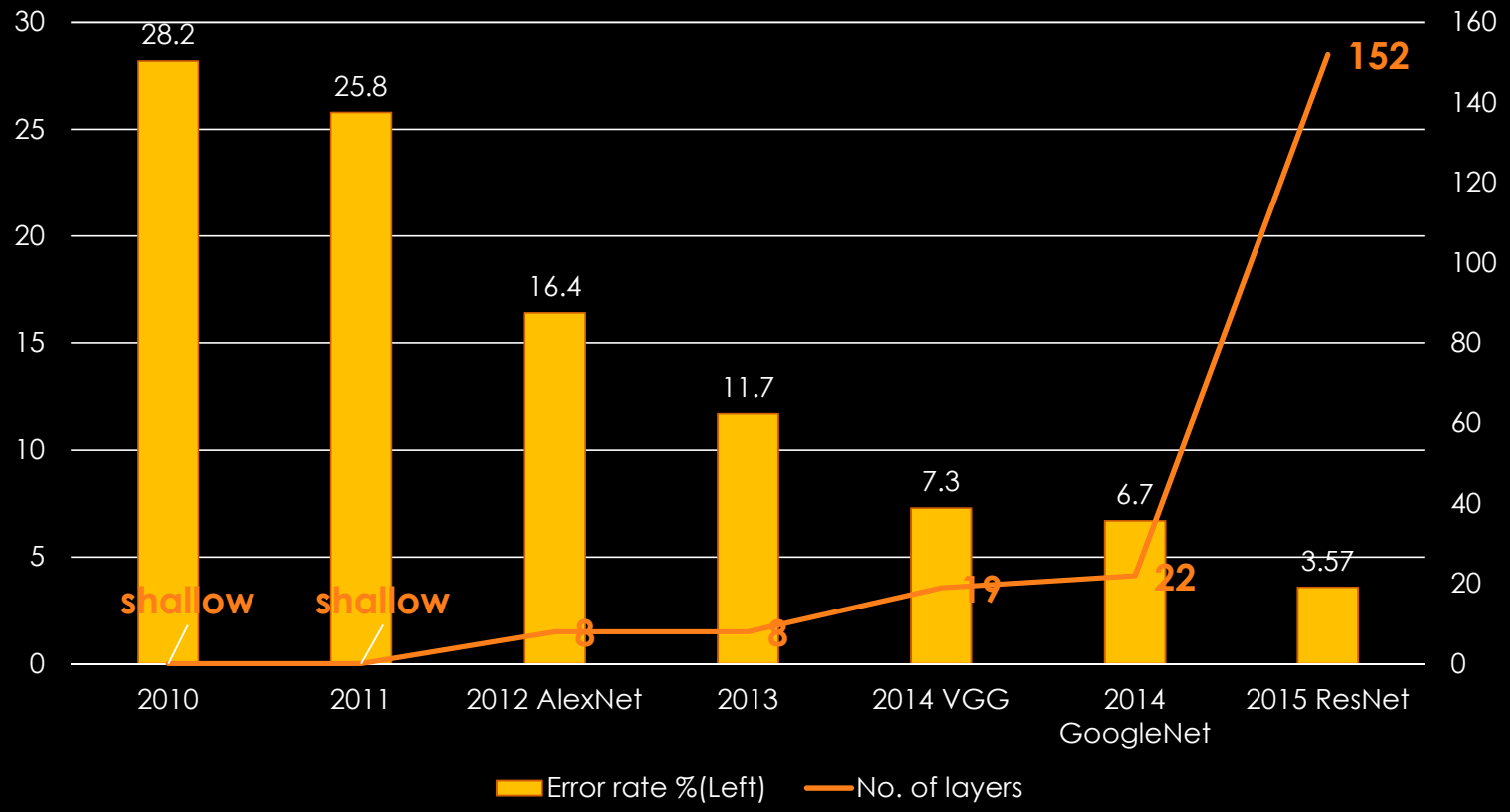
- A picture is only a group of pixels for a computer
- Modern AI nets learn features of objects

Images source: CC dataset

OBJECT CLASSIFICATION

- Modern AI algorithms for object classification
  - AlexNet, 5 CNN layers and 3 FC layers, 2012
  - VGG, 16 CNN layers and 3 FC layers, 2014
  - GoogLenet, 21 CNN layers and 1 FC layer, 2014
  - ResNet, 151 CNN layers and 1 FC layer, 2015
- Foundation of object detection
- Limitation
  - One object in one picture, no localization

ILSVRC(IMAGENET LARGE SCALE VISUAL RECOGNITION COMPETITION)

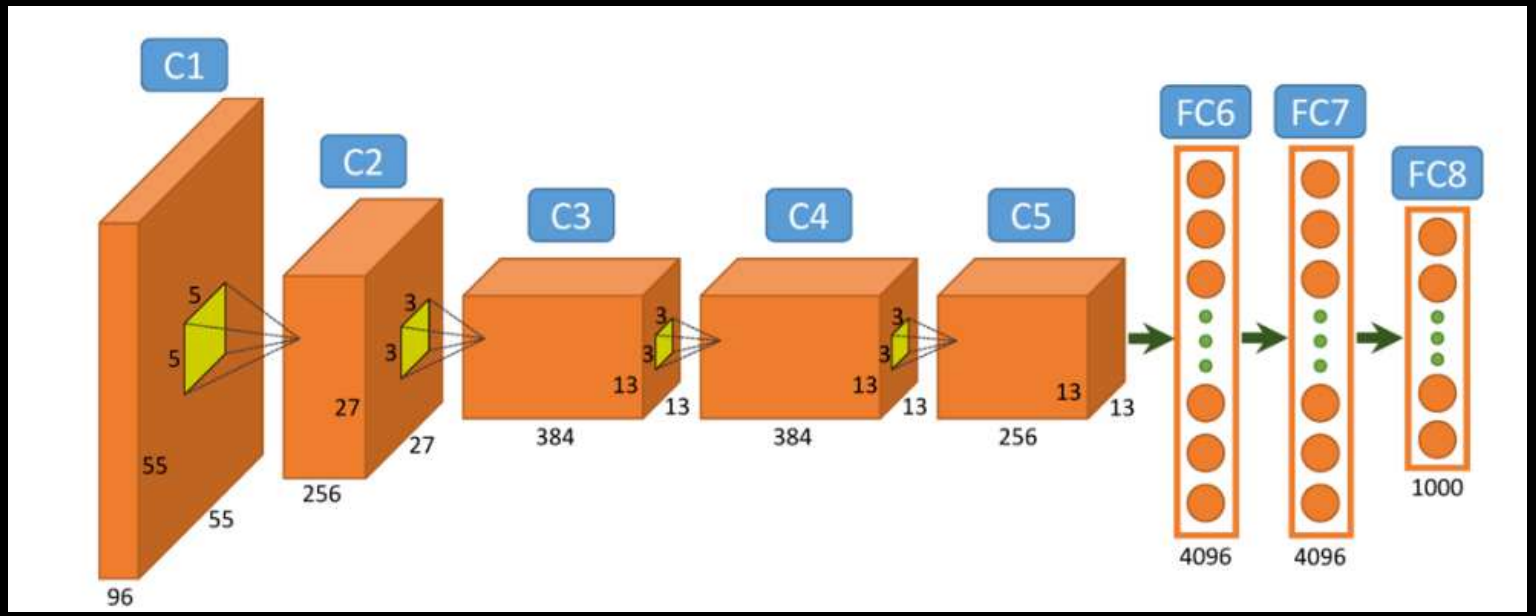


OBJECT CLASSIFICATION

- Modern AI algorithms for Object Classification
  - AlexNet, 5 CNN layers and 3 FC layers, 2012
  - VGG, 16 CNN layers and 3 FC layers, 2014
  - GoogLenet, 21 CNN layers and 1 FC layer, 2014
  - ResNet, 151 CNN layers and 1 FC layer, 2015
- Foundation of object detection
- Limitation
  - One object in one picture, no localization

ALEXNET

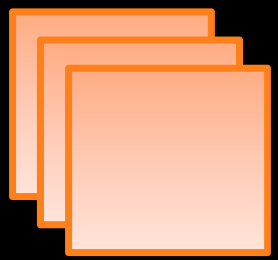
- CONV Layers: 5
- Fully Connected Layers: 3
- Weights: 61M
- MACs: 724M



ALEXNET

Layers	Filter Size (R x S)	Filters (M)	Channels (C)	Stride
1	11x11	96	3	4
2	5x5	256	48	1
3	3x3	384	256	1
4	3x3	384	192	1
5	3x3	256	192	1

Layer 1



34k Params.
105M MACs

Layer 2

307k Params.
224M MACs

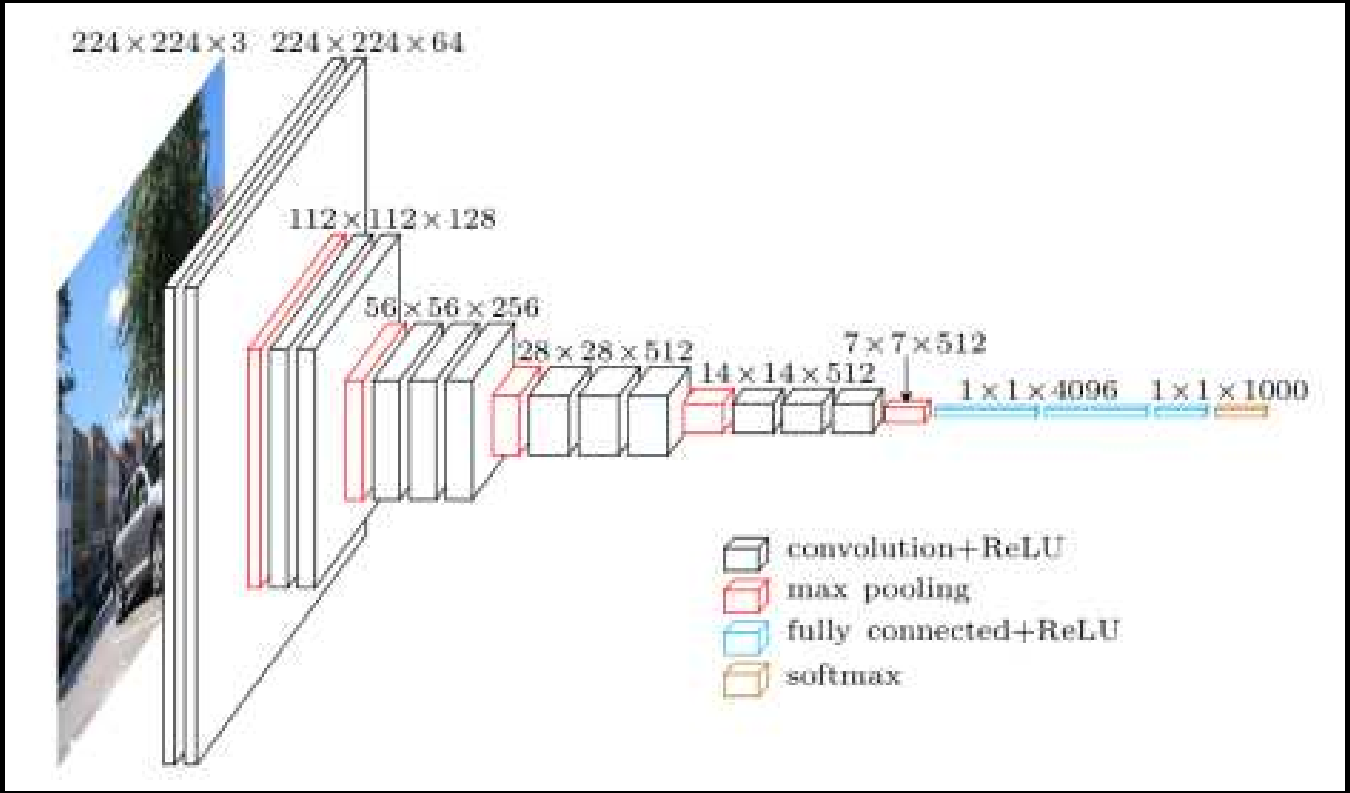
Layer 3

885k Params.
150M MACs

OBJECT CLASSIFICATION

- Modern AI algorithms for Object Classification
  - AlexNet, 5 CNN layers and 3 FC layers, 2012
  - VGG, 16 CNN layers and 3 FC layers, 2014
  - GoogLenet, 21 CNN layers and 1 FC layer, 2014
  - ResNet, 151 CNN layers and 1 FC layer, 2015
- Foundation of object detection
- Limitation
  - One object in one picture, no localization

VGG



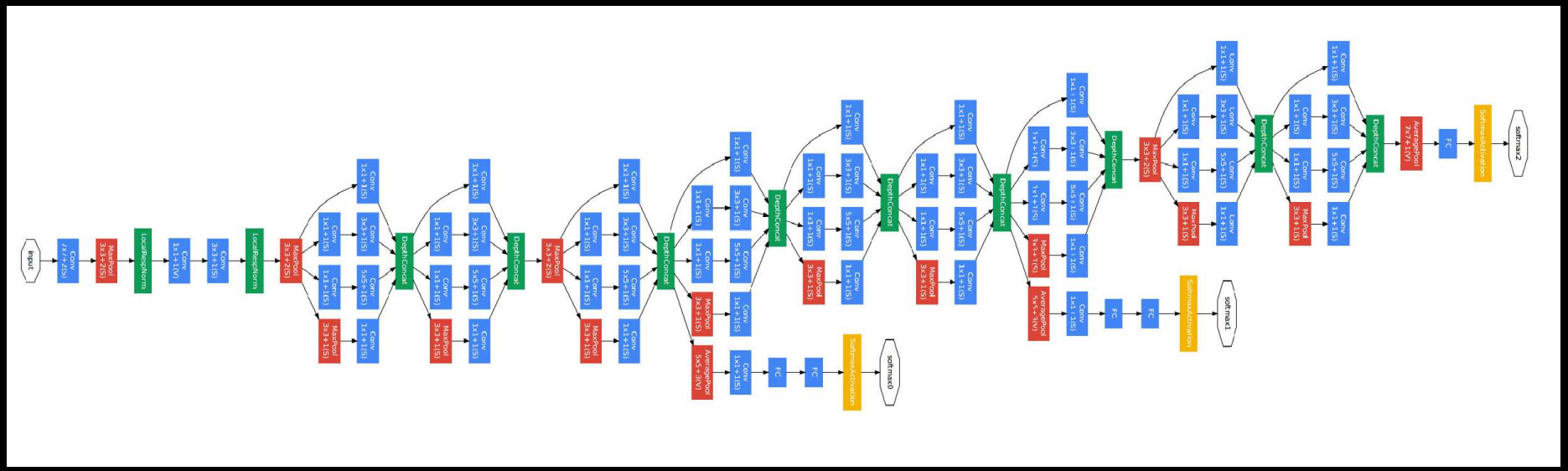
- CONV Layers: 16
- Fully Connected Layers: 3
- Weights: 138M
- MACs: 15.5G

OBJECT CLASSIFICATION

- Modern AI algorithms for Object Classification
  - AlexNet, 5 CNN layers and 3 FC layers, 2012
  - VGG, 16 CNN layers and 3 FC layers, 2014
  - GoogLenet, 21 CNN layers and 1 FC layer, 2014
  - ResNet, 151 CNN layers and 1 FC layer, 2015
- Foundation of object detection
- Limitation
  - One object in one picture, no localization

GOOGLENET

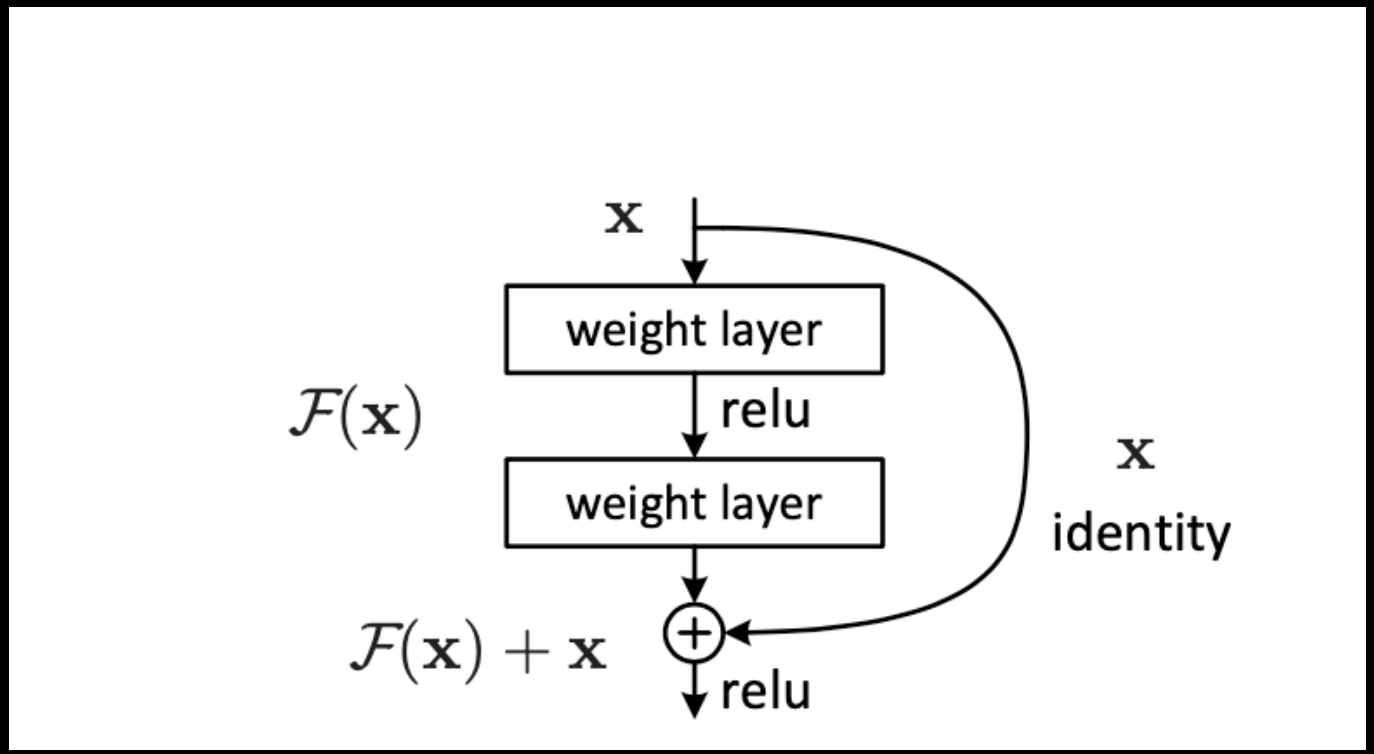
- CONV Layers: 21
- Fully Connected Layers: 1
- Weights: 7.0M
- MACs: 1.43G



OBJECT CLASSIFICATION

- Modern AI algorithms for Object Classification
  - AlexNet, 5 CNN layers and 3 FC layers, 2012
  - VGG, 16 CNN layers and 3 FC layers, 2014
  - GoogLenet, 21 CNN layers and 1 FC layer, 2014
  - ResNet, 151 CNN layers and 1 FC layer, 2015
- Foundation of object detection
- Limitation
  - One object in one picture, no localization

RESNET



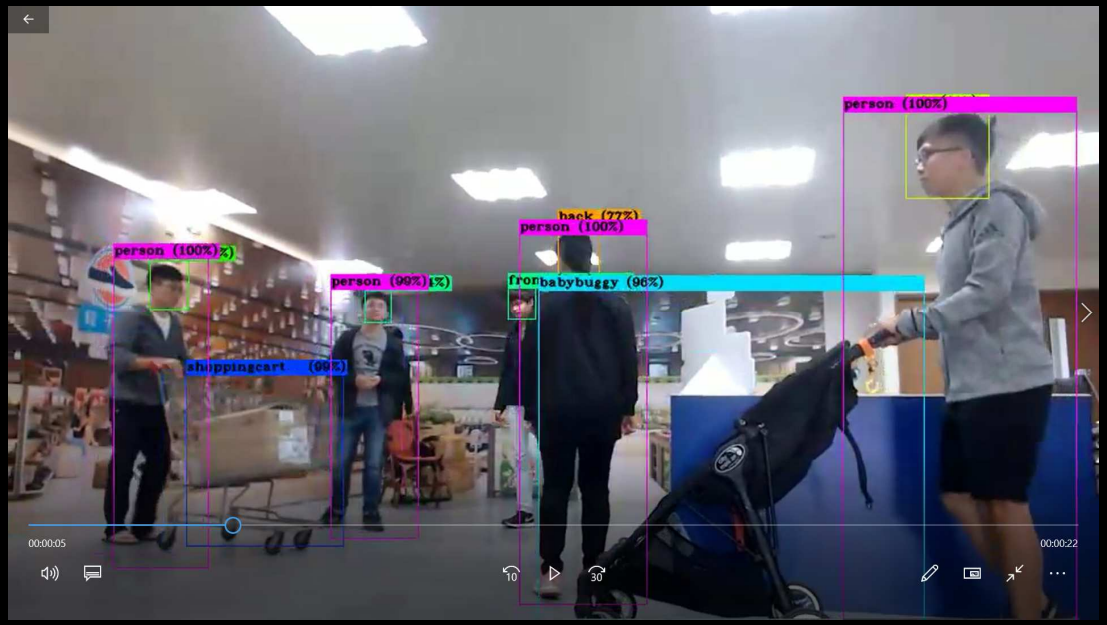
- Main idea
  - Residual layer
- CONV Layers: 151
- Fully Connected Layers: 1
- Weights: 25.5M
- MACs: 3.9G

OUTLINE

- What are AI visual algorithms?
  - How does a computer classify pictures?
  - How does a computer detect objects?
  - What else can AI visual algorithms do?
- How to evaluate performance of AI visual algorithms?

HOW DOES A COMPUTER DETECT OBJECTS?

- Besides class, the computer needs to know the location of each object.



Images source: FCU SoC Lab

OBJECT DETECTION

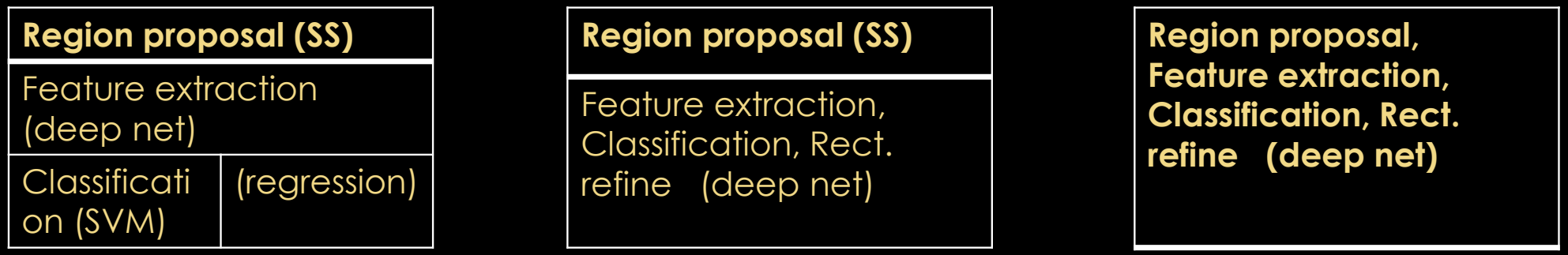
- Modern AI algorithms for Object Detection
  - RCNN (Region-based CNN), fast RCNN, faster RCNN
  - YOLO (You Only Look Once)
  - SSD (Single Shot Detection)

OBJECT DETECTION

- Modern AI algorithms for Object Detection
  - RCNN (Region-based CNN), fast RCNN, faster RCNN
  - YOLO (You Only Look Once)
  - SSD (Single Shot Detection)

RCNN (REGION-BASED CNN), FAST RCNN, FASTER RCNN

Two-stage ways

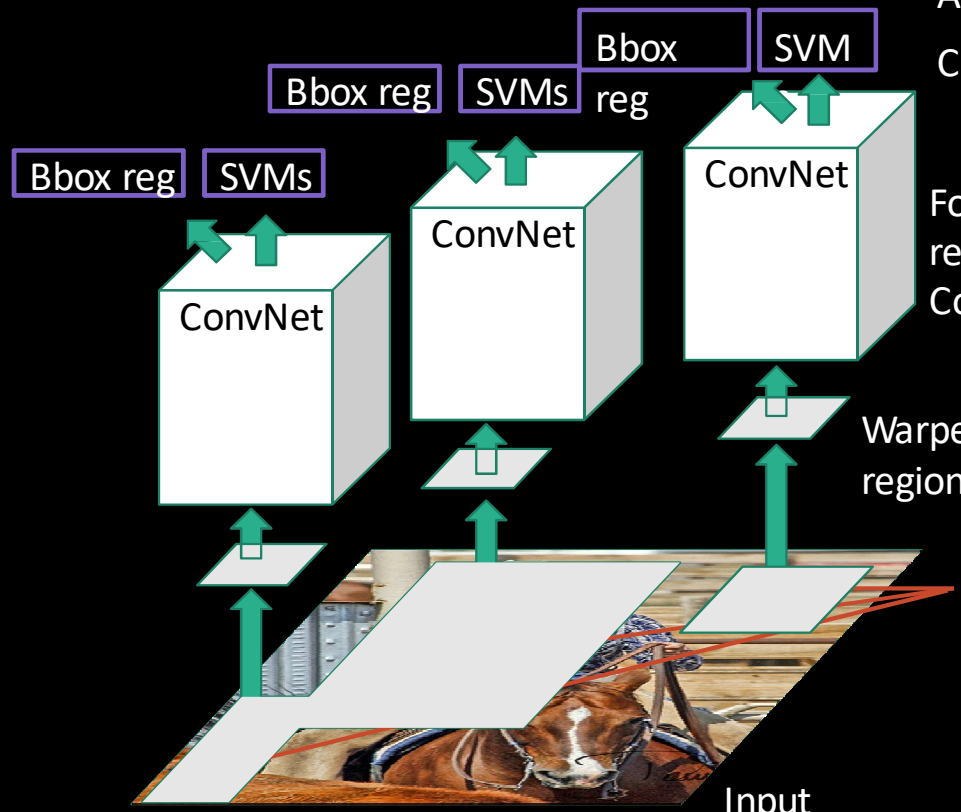


RCNN
Slow in both training and testing

Fast-RCNN
Few seconds per frame

Faster-RCNN
A dozen of fps on k40

TWO-STAGE WAYS



- Region proposal
- Conv. layers

Girshick et al.

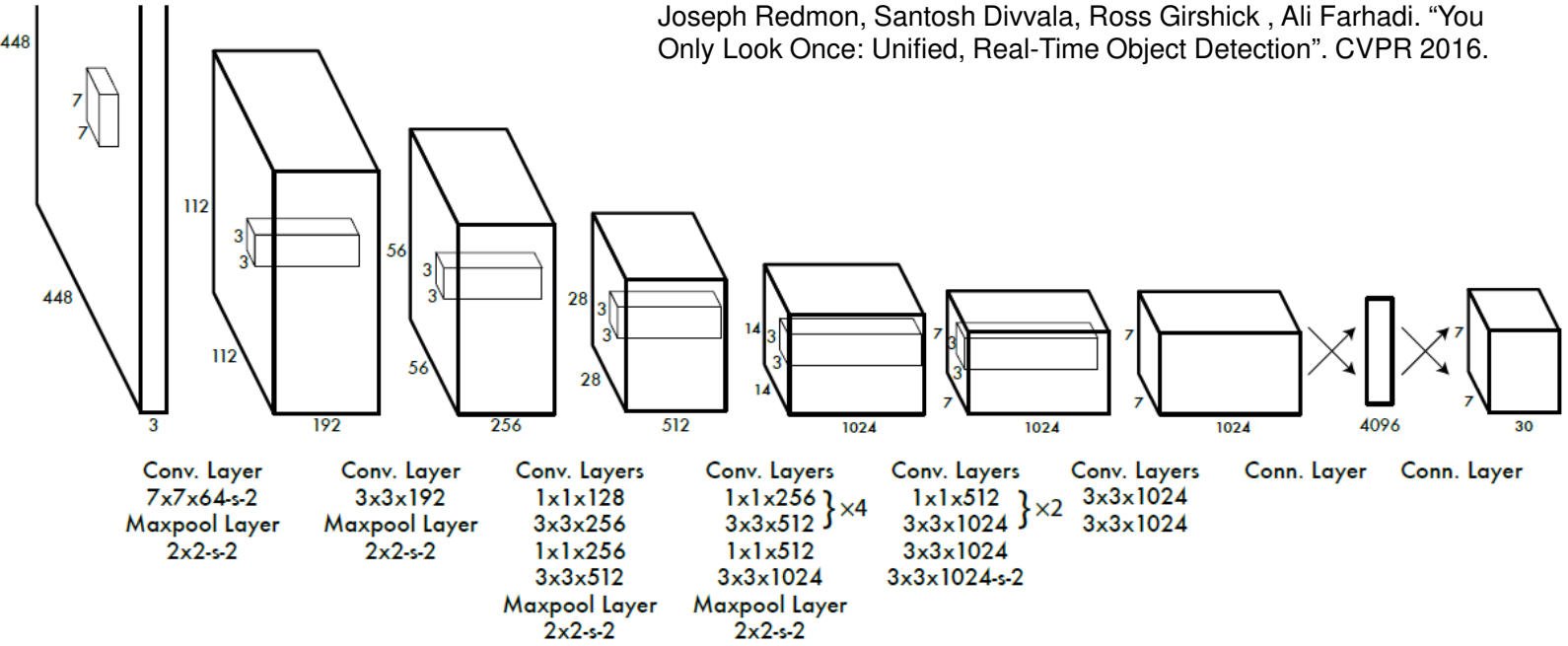
OBJECT DETECTION

- Modern AI algorithms for Object Detection
  - RCNN (Region-based CNN), fast RCNN, faster RCNN
  - YOLO (You Only Look Once)
  - SSD (Single Shot Detection)

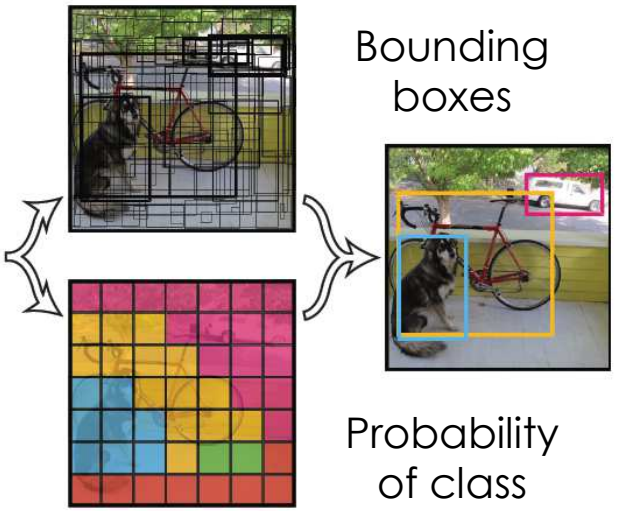
YOLO (YOU ONLY LOOK ONCE)

- One-stage way

Joseph Redmon, Santosh Divvala, Ross Girshick , Ali Farhadi. “You Only Look Once: Unified, Real-Time Object Detection”. CVPR 2016.



Classification and localization



YOLO V3

YOLO v1 (2015)

YOLO v2 (2017)

YOLO v3 (2018)

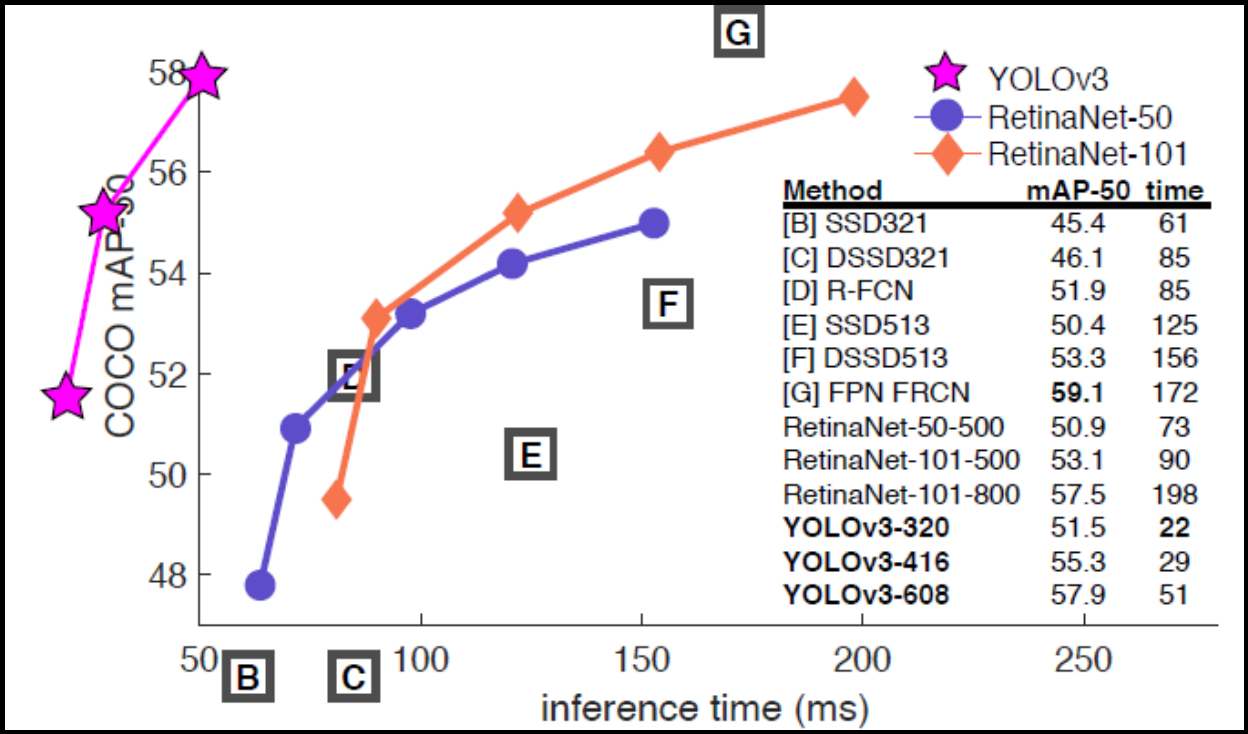
BN,

ResNet,

Multiscale

https://pjreddie.com/darknet/yolo/

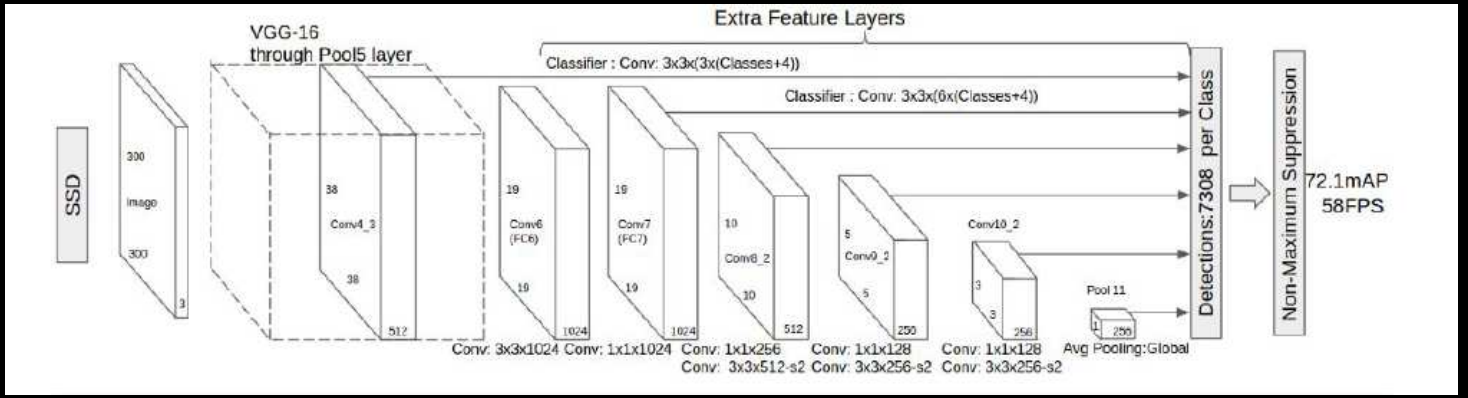
Fast and accurate under mAP-50



Method	mAP-50	time
[B] SSD321	45.4	61
[C] DSSD321	46.1	85
[D] R-FCN	51.9	85
[E] SSD513	50.4	125
[F] DSSD513	53.3	156
[G] FPN FRCN	**59.1**	172
RetinaNet-50-500	50.9	73
RetinaNet-101-500	53.1	90
RetinaNet-101-800	57.5	198
YOLOv3-320	51.5	**22**
YOLOv3-416	55.3	29
YOLOv3-608	57.9	51

OBJECT DETECTION

- Modern AI algorithms for Object Detection
  - RCNN (Region-based CNN), fast RCNN, faster RCNN
  - YOLO (You Only Look Once)
  - SSD (Single Shot Detection)
    - +Multi-scale feature maps
    - - FC layers



OUTLINE

- What are AI visual algorithms?
  - How does a computer classify pictures?
  - How does a computer detect objects?
  - What else can AI visual algorithms do?
- How to evaluate performance of AI visual algorithms?

INTRODUCTION OF AI VISUAL ALGORITHMS

- Evolution of functionalities
  - Classification, Detection, Segmentation (Spatial domain)
  - Behavior Analysis (Spatial and temporal domains)
  - Text-to-image, image-to-text, image synthesis (GAN)
  - Joint object detection and segmentation
  - Joint object and action detection
- Evolution of implementation
  - 32-bit floating-point → 16-bit fixed-point → few-bit fixed-point

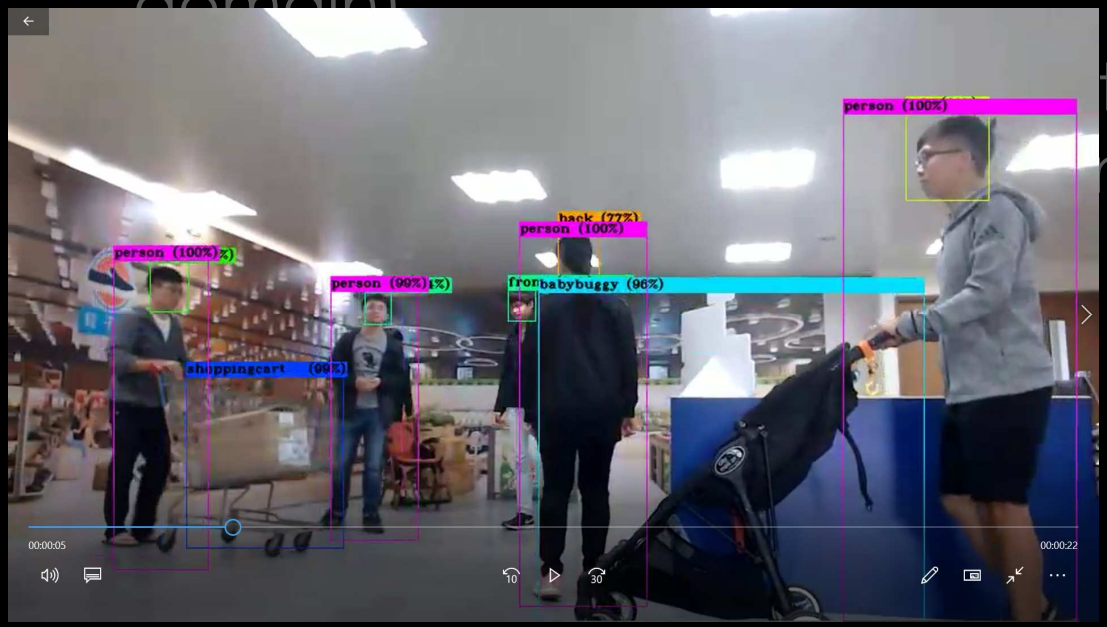
INTRODUCTION OF AI VISUAL ALGORITHMS

- Evolution of Functionalities
  - Classification → Detection → Segmentation (Spatial domain)

Images source: CC dataset

INTRODUCTION OF AI VISUAL ALGORITHMS

- Evolution of Functionalities
  - Classification → Detection → Segmentation (Spatial domain)
    temporal domains)
    mage synthesis (GAN)



Images source: FCU SoC Lab

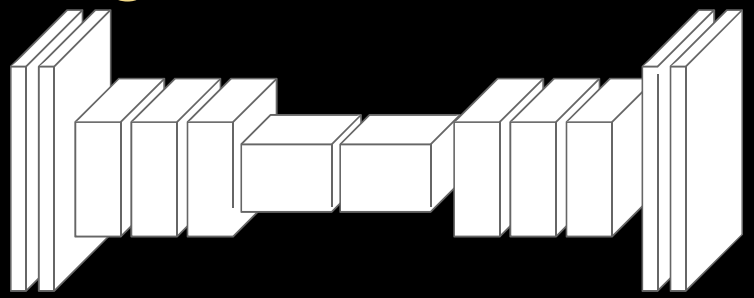
INTRODUCTION OF AI VISUAL ALGORITHMS

- Evo es
  - → Segmentation (Spatial
  - and temporal domains)
  - ext, image synthesis (GAN)

Images source: coco dataset

SEMANTIC SEGMENTATION

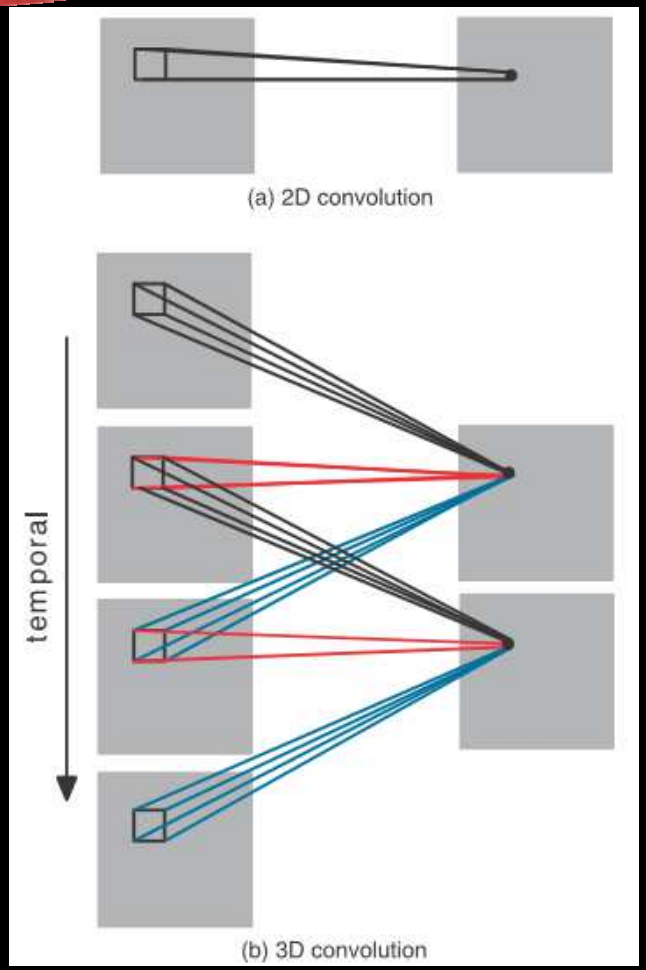
- Common Semantic Segmentation Styles
  - Downsampling path: extracts coarse features
  - Upsampling path: recovers input image resolution
  - Multi-scale detection
  - Post-processing (optional): refines predictions (CRF, condition random field)
- Modern AI algorithms for Semantic Segmentation
  - DeepLab v1, v2, v3, v3+ (ICLR15)
  - Dilated convolutions 2016



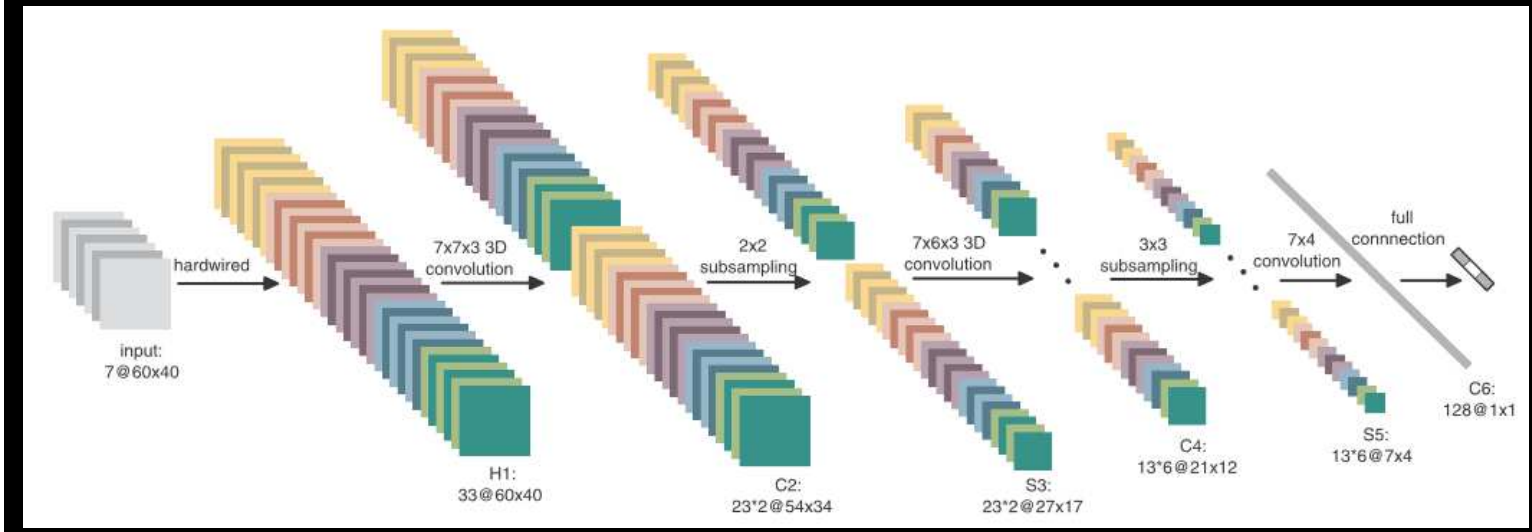
INTRODUCTION OF AI VISUAL ALGORITHMS

- Evolution of Functionalities
  - Classification → Detection → Segmentation (Spatial domain)
  - Behavior Analysis (Spatial and temporal domains)
  - Text-to-image, image-to-text, image synthesis (GAN)

3D CNN



- To compute features from both spatial and temporal dimensions
  - Input 7 frames



Source: 3D Convolutional Neural Networks for Human Action Recognition, TPAMI, 2013.

INTRODUCTION OF AI VISUAL ALGORITHMS

- Evolution of Functionalities
  - Classification → Detection → Segmentation (Spatial domain)
  - Behavior Analysis (Spatial and temporal domains)
  - Text-to-image, image-to-text, image synthesis (GAN)

INTRODUCTION OF AI VISUAL ALGORITHMS

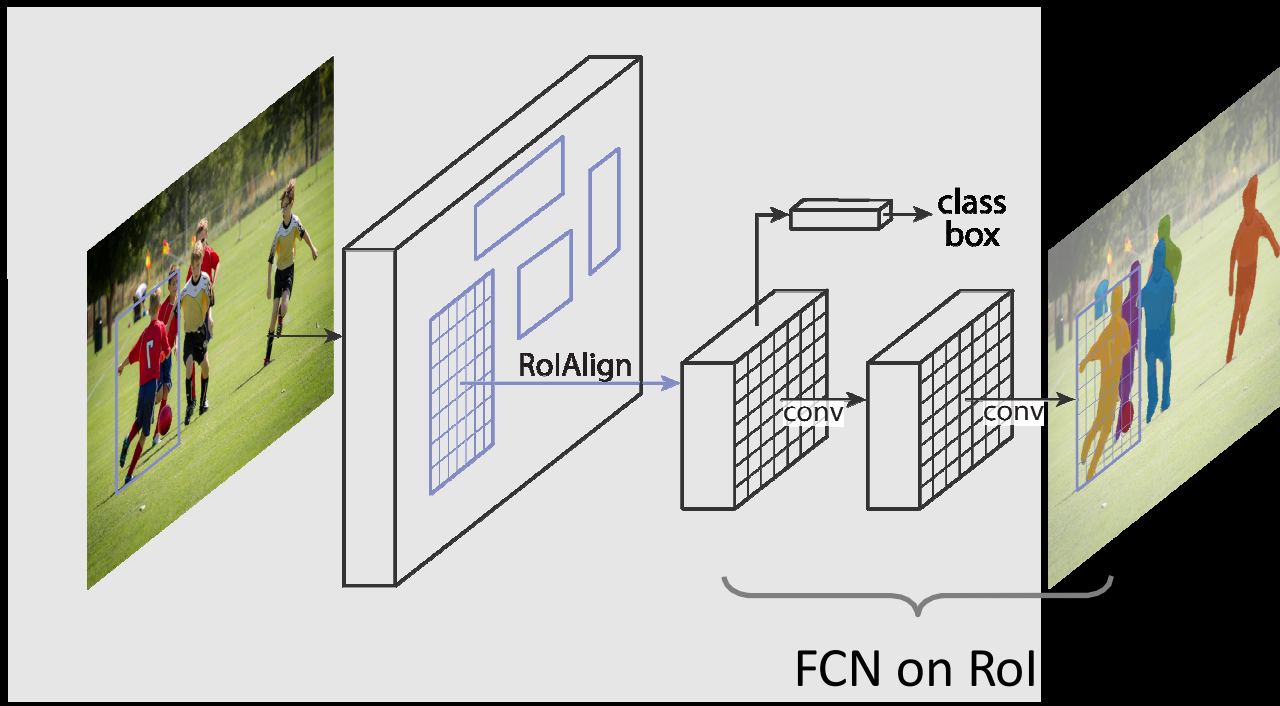
- Evo ionalities
  - C etectio
    o
  - B (Spatia
  - T age-to-

INTRODUCTION OF AI VISUAL ALGORITHMS

- Evolution of functionalities
  - Classification → Detection → Segmentation (Spatial domain)
  - Behavior Analysis (Spatial and temporal domains)
  - Text-to-image, image-to-text, image synthesis (GAN)
  - Joint object detection and segmentation
  - Joint object and action detection
- Evolution of implementation
  - 32-bit floating-point → 16-bit fixed-point → few-bit fixed-point

Joint object detection and segmentation

- Instance segmentation
- Mask R-CNN = **Faster R-CNN** with **FCN** on RoIs
  - A binary mask that says whether or not a given pixel is part of an object
  - Nearly as fast as faster R-CNN



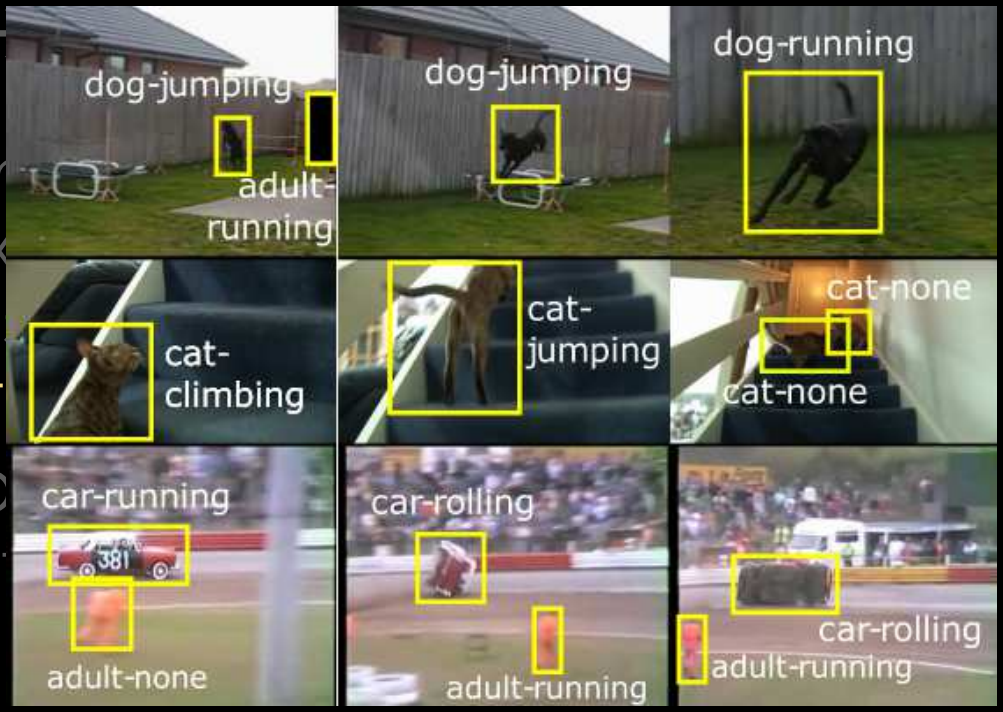
Source:2017, mask R-CNN

INTRODUCTION OF AI VISUAL ALGORITHMS

- Evolution of functionalities
  - Classification → Detection → Segmentation (Spatial domain)
  - Behavior Analysis (Spatial and temporal domains)
  - Text-to-image, image-to-text, image synthesis (GAN)
  - Joint object detection and segmentation
  - Joint object and action detection
- Evolution of implementation
  - 32-bit floating-point → 16-bit fixed-point → few-bit fixed-point

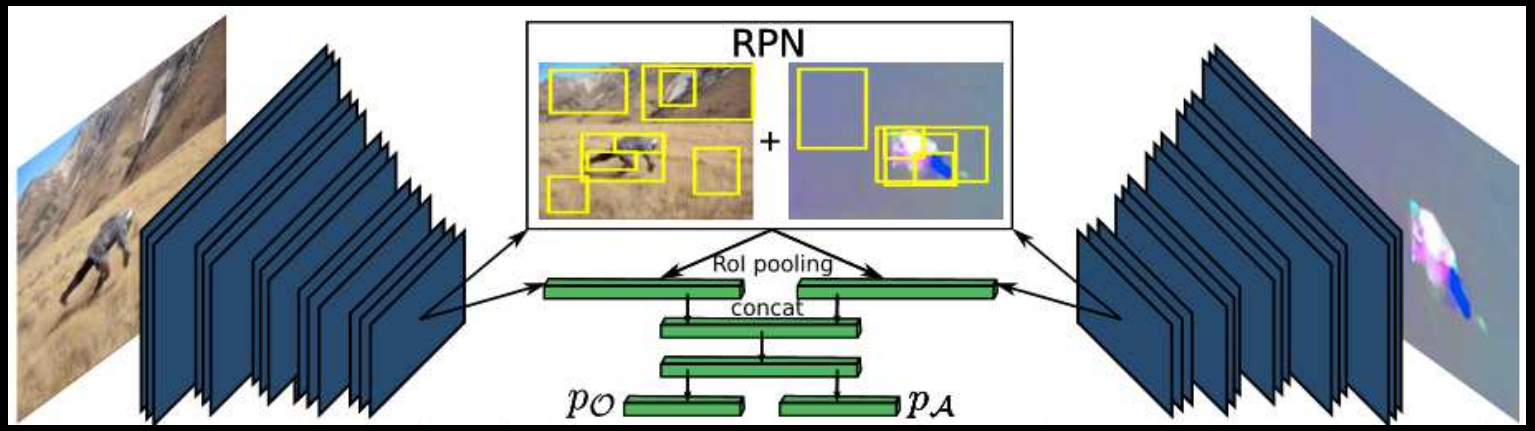
INTRODUCTION OF AI VISUAL ALGORITHMS

- Evolution of functionalities
  - Classification → Detection → domain)
  - Behavior Analysis (Spatial ar
  - Text-to-image, image-to-tex
  - Joint object detection and
  - Joint object and action det
- Evolution of implementatic
  - 32-bit floating-point → 16-bi point



Images source: Joint learning of object and action detectors, ICCV 2019

MULTITASK NETWORK FOR JOINT OBJECT-ACTION DETECTION



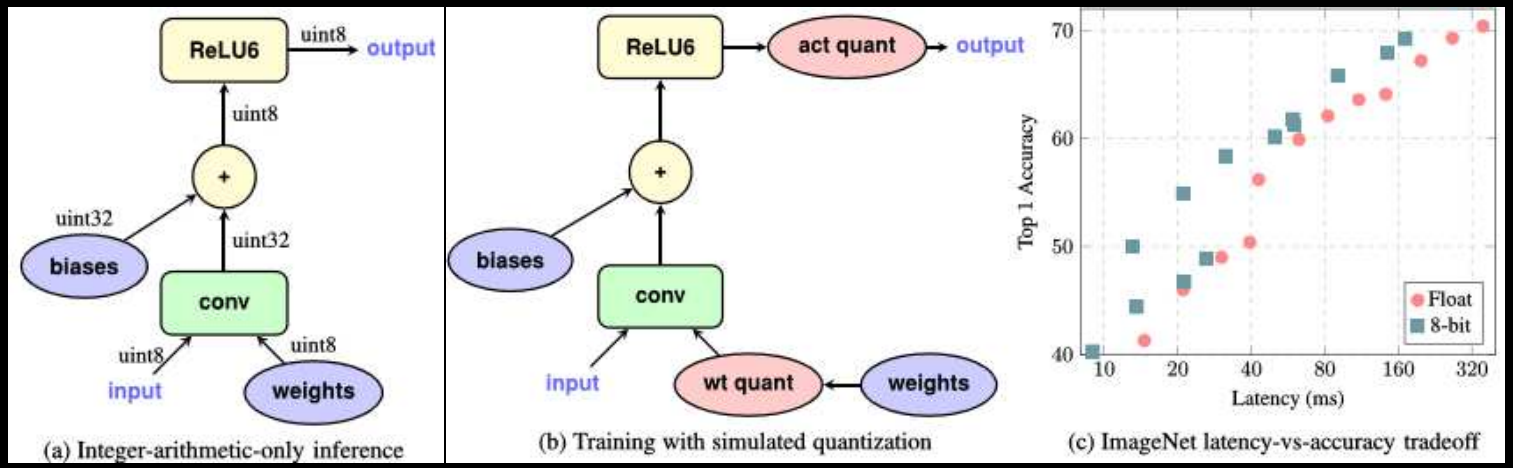
- Rely on Faster RCNN and its two-stream variant
  - Two streams: appearance and motion
- A Region Proposal Network (RPN) extracts candidate bounding boxes independently for each stream
- A Region-of-Interest (RoI) pooling layer uses the set union of the two RPNs and aggregates features for each candidate box

Source: Joint learning of object and action detectors, ICCV 2019

INTRODUCTION OF AI VISUAL ALGORITHMS

- Evolution of functionalities
  - Classification → Detection → Segmentation (Spatial domain)
  - Behavior Analysis (Spatial and temporal domains)
  - Text-to-image, image-to-text, image synthesis (GAN)
  - Joint object detection and segmentation
  - Joint object and action detection
- Evolution of implementation
  - 32-bit floating-point → 16-bit fixed-point → few-bit fixed-point

QUANTIZATION AND TRAINING OF NEURAL NETWORKS



- Quantization scheme
  - $r = S(q - Z)$
  - r: real value; S: scale; q: quantized value; Z: zero point

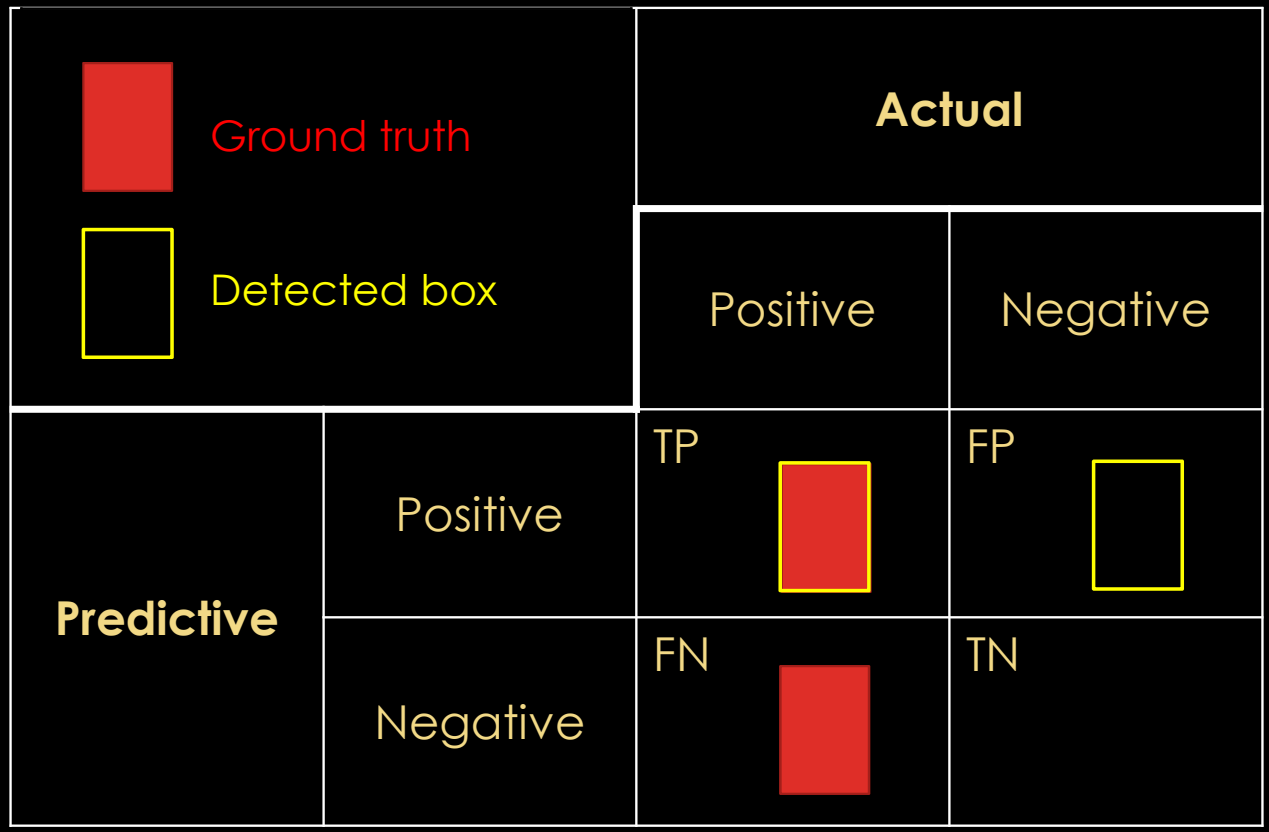
Source: Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference, CVPR 2018.

PERFORMANCE EVALUATION INDEXES

- TP, FP, TN, FN
- Precision, Recall
- mAP (mean Average Precision)
- Important Parameters

PERFORMANCE EVALUATION INDEXES

- TP, FP, TN, FN
  - TP: True Positive
  - FP: False Positive
  - TN: True Negative
  - FN: False Negative

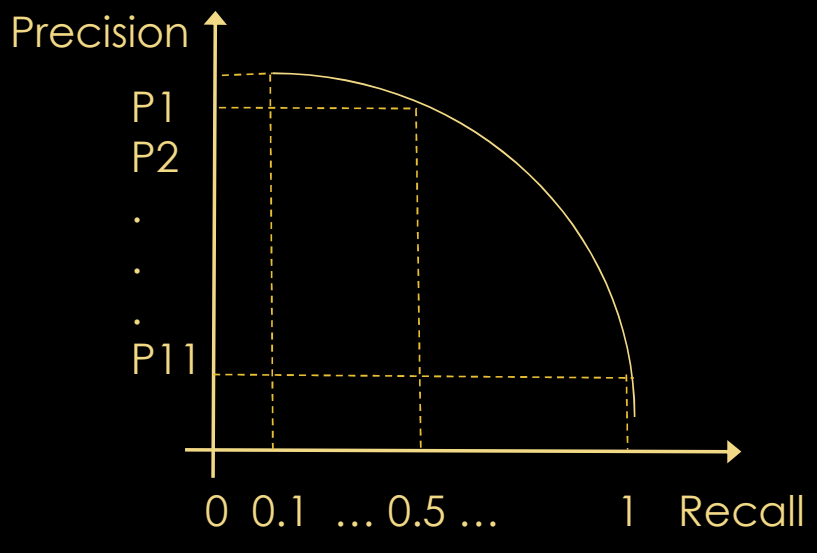


PERFORMANCE EVALUATION INDEXES

- Precision, Recall

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

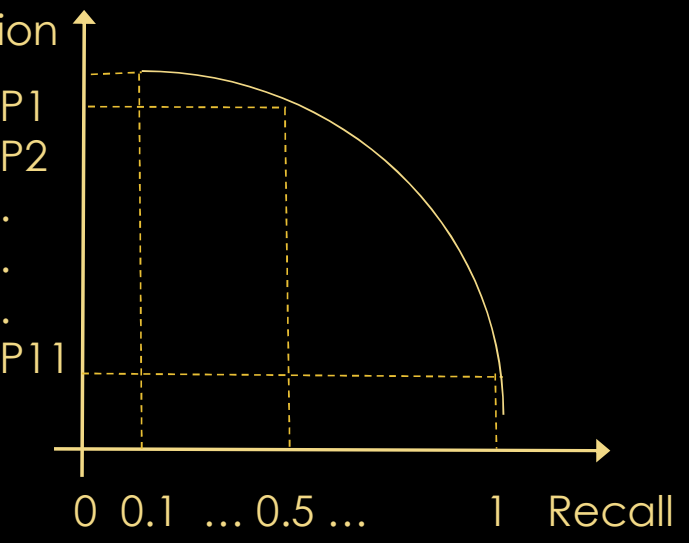


PERFORMANCE EVALUATION INDEXES

- mAP (mean Average Precision)
  - AP: the average precision of precisions of different recalls
  - mAP: the mean of APs of different kinds of objects

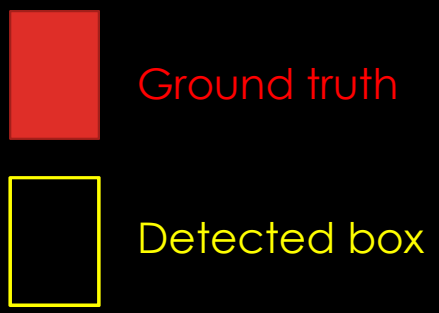
$$AP = \int_0^1 p(r)\,\mathrm{d}r$$

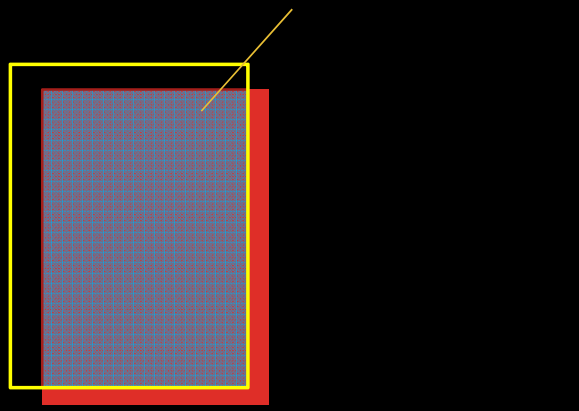
$$mAP = \frac{\sum_{i=1}^{n} AP_i}{n}$$



PERFORMANCE EVALUATION INDEXES

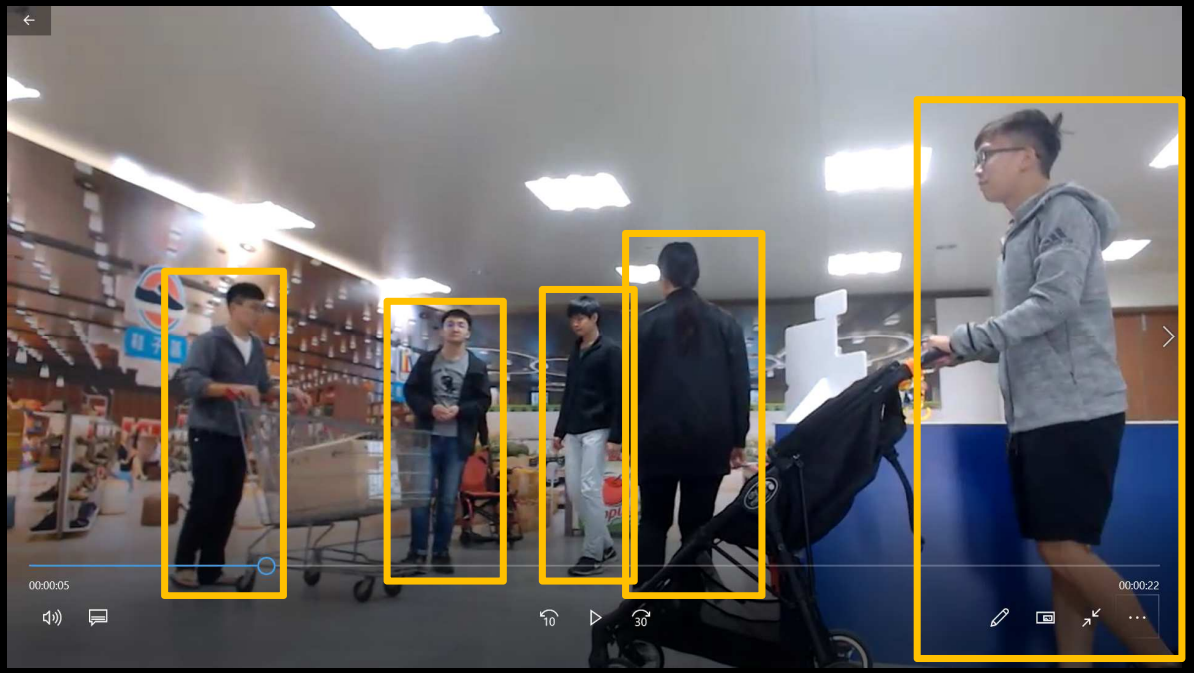
- Important Parameters
  - IoU (Intersection over Union)
  - confidence threshold





EXAMPLE 1/4

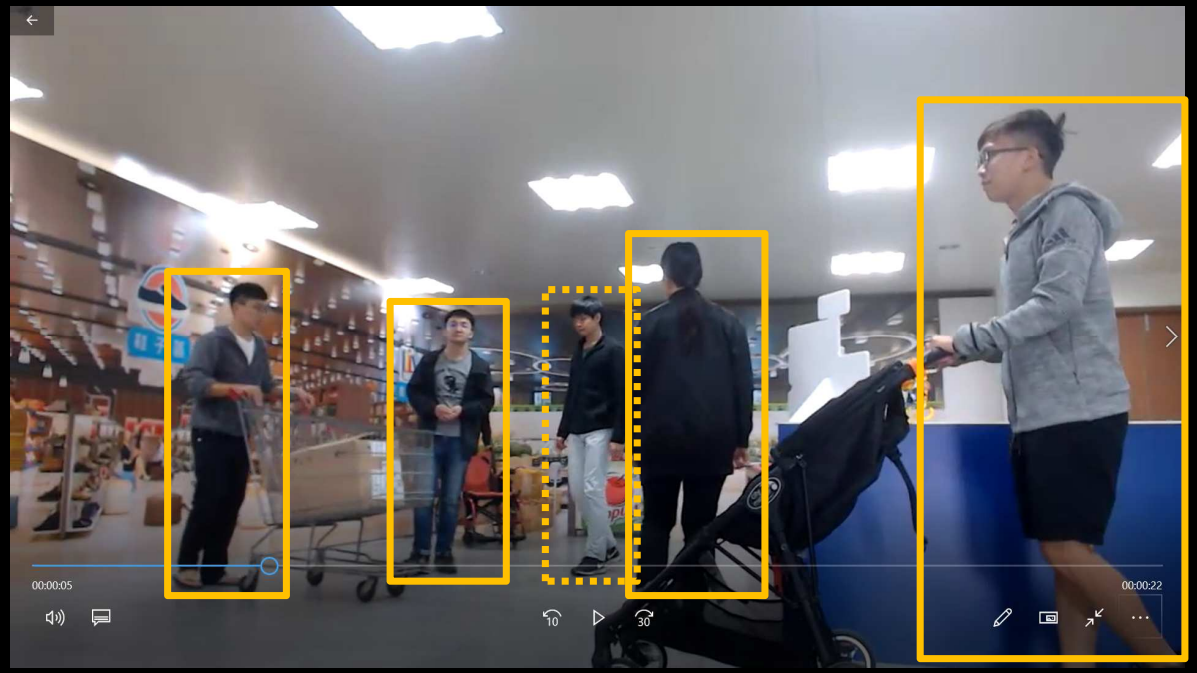
- Class: people
  - TP: 5
  - FP: 0
  - FN: 0
  - Precision: 5/5
  - Recall: 5/5



EXAMPLE 2/4

- Class: people
  - TP: 4
  - FP: 0
  - FN: 1
  - Precision: 4/4
  - Recall: 4/5

		Actual	
		Positive	Negative
Predictive	Positive	TP	FP
	Negative	FN	TN



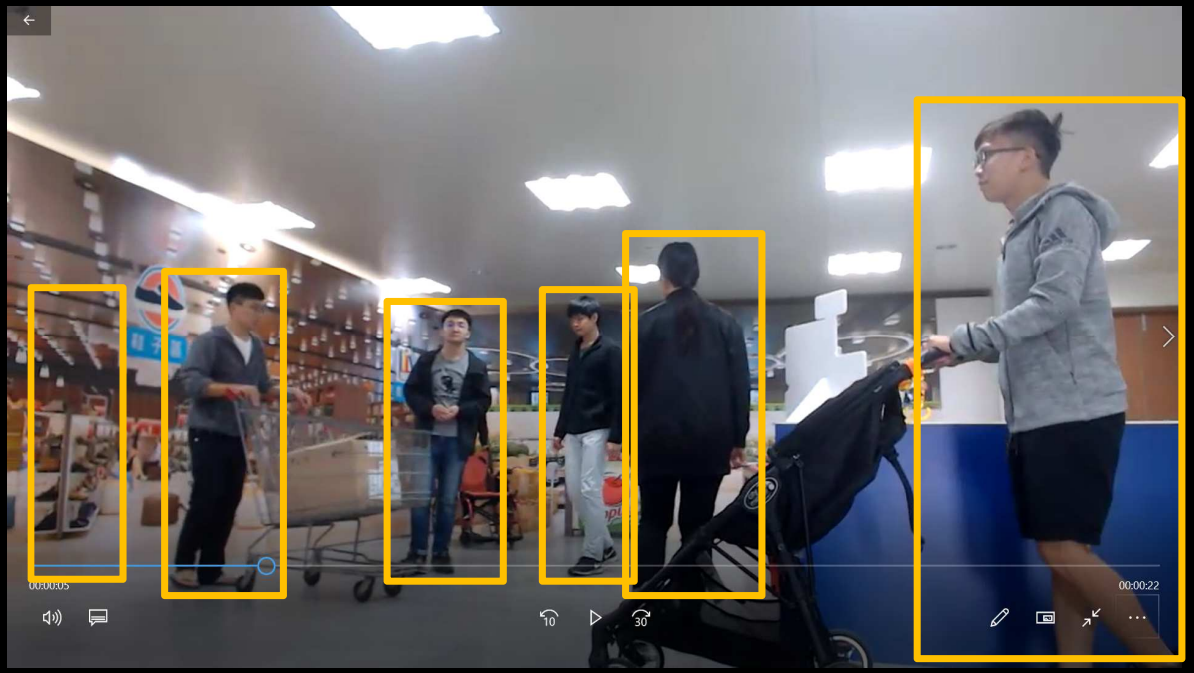
$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

EXAMPLE 3/4

- Class: people
  - TP: 5
  - FP: 1
  - FN: 0
  - Precision: 5/6
  - Recall: 5/5

		Actual	
		Positive	Negative
Predictive	Positive	TP	FP
	Negative	FN	TN



$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

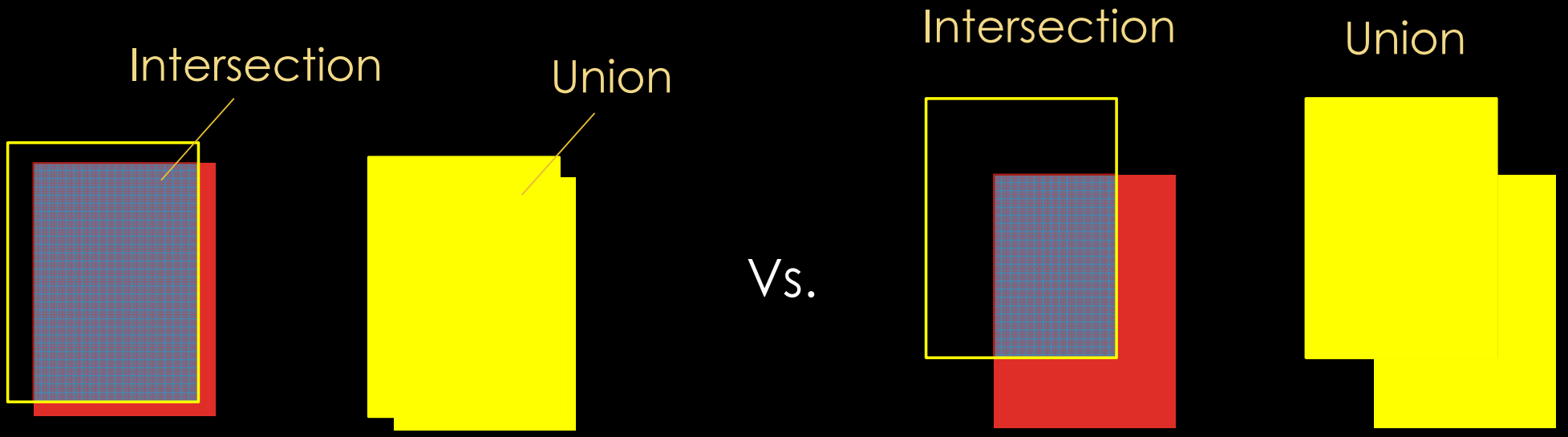
EXAMPLE 4/4

- Assume
  - Recall 1.0 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0.0
  - Precision 0.70 0.74 0.78 0.82 0.85 0.89 0.93 0.96 0.98 0.99 1.00
  - AP=(0.7+0.74+…+1)/11=0.88

$$mAP = \frac{\sum_{i=1}^{n} AP_i}{n}$$ for n classes

NOTE 1/2

- How does IoU affect AP?
  - Judging criteria of a nice shot



Case I: IoU=0.88

Case II: IoU=0.58

NOTE 2/2

- Commonly used indexes
  - AP-50: IoU=0.5 as the threshold
    - Both case I (IoU=0.88) and case II (IoU=0.58) get 1 TP
  - AP-75: IoU=0.75 as the threshold
    - Case I (IoU=0.88) is TP, but case II (IoU=0.58) is not
    - Besides losing 1 TP, case II generates 1 FP and 1 FN simultaneously
  - AP@[0.5 : 0.95]: from IoU=0.5 to IoU=0.95 with a step size of 0.05 (adopted in COCO dataset)

THANK YOU FOR LISTENING

OUTLINE

- What are AI visual algorithms?
- How to evaluate performance of AI visual algorithms?