You Only Look Once: Unified, Real-Time Object Detection

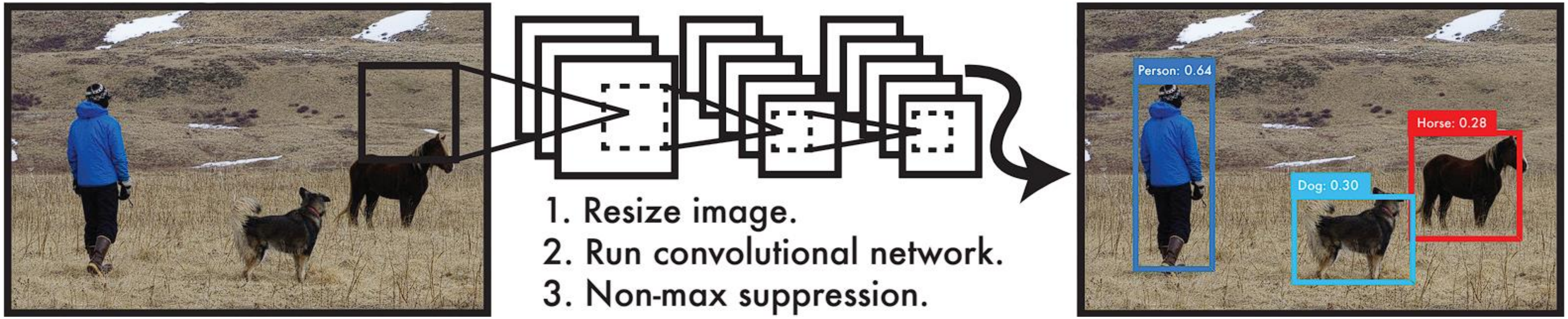
BURAK BOZDAĞ - 504211552

Introduction

- Humans glance at an image and instantly know what it is
- YOLO: Fast, generalizable, maintains accuracy

Introduction

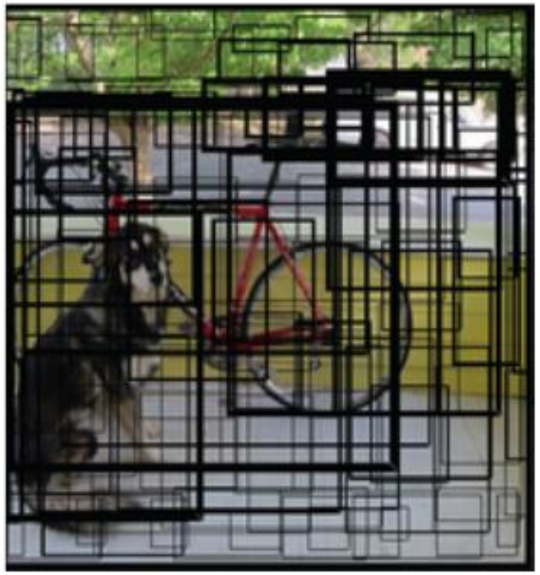
- Resizing to 448 x 448
- Running single CNN
- Thresholding by the model's confidence



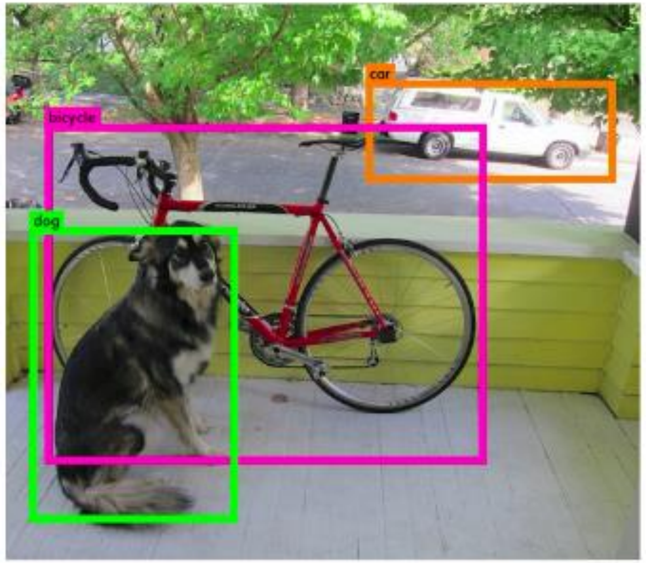
1. Resize image.
2. Run convolutional network.
3. Non-max suppression.

Non-Maximal Suppression

- Intersection Over Union (IOU) = Area of Overlap / Area of Union
- Selecting the right bounding box
- Eliminating redundant ones



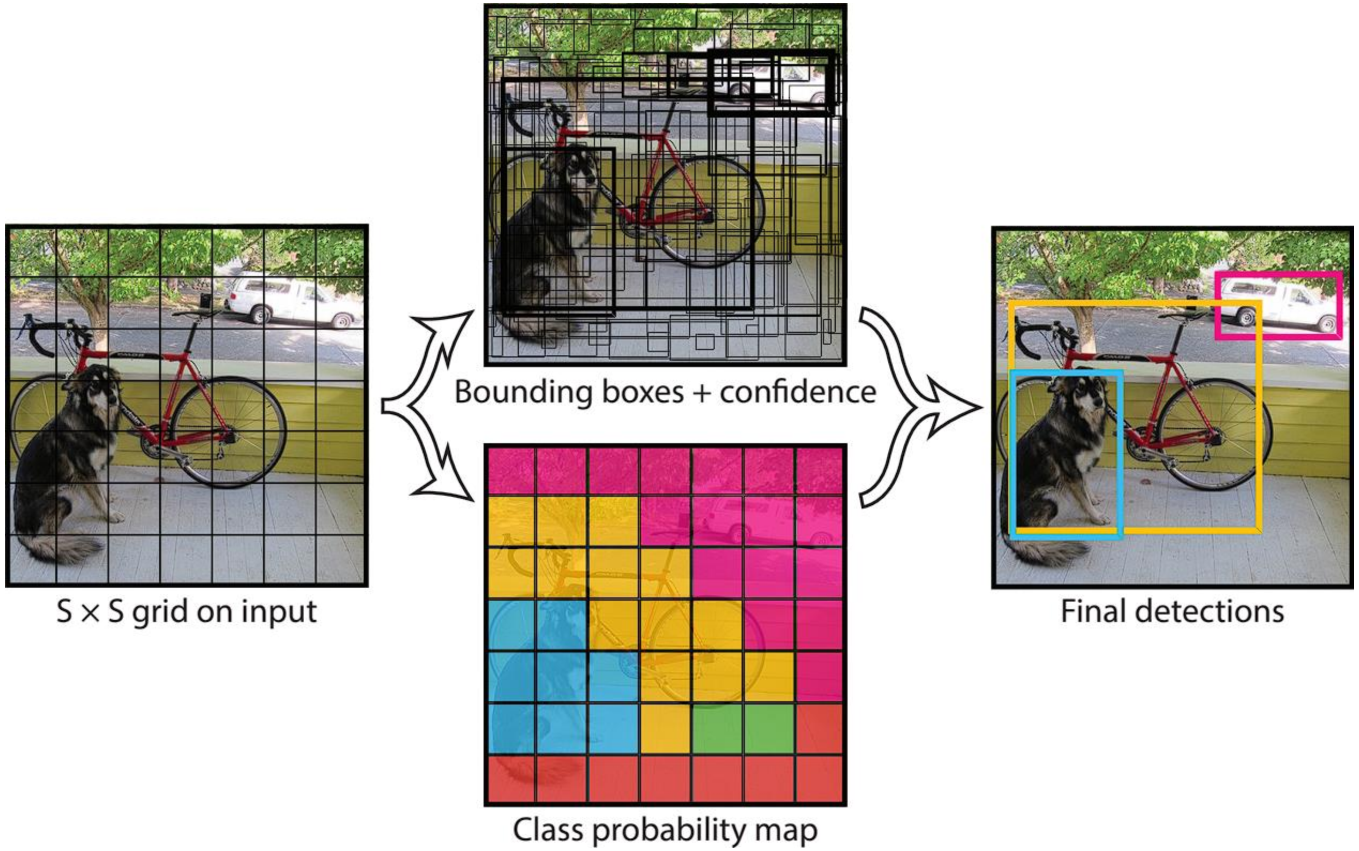
Multiple Bounding Boxes



Final Bounding Boxes

Unified Detection

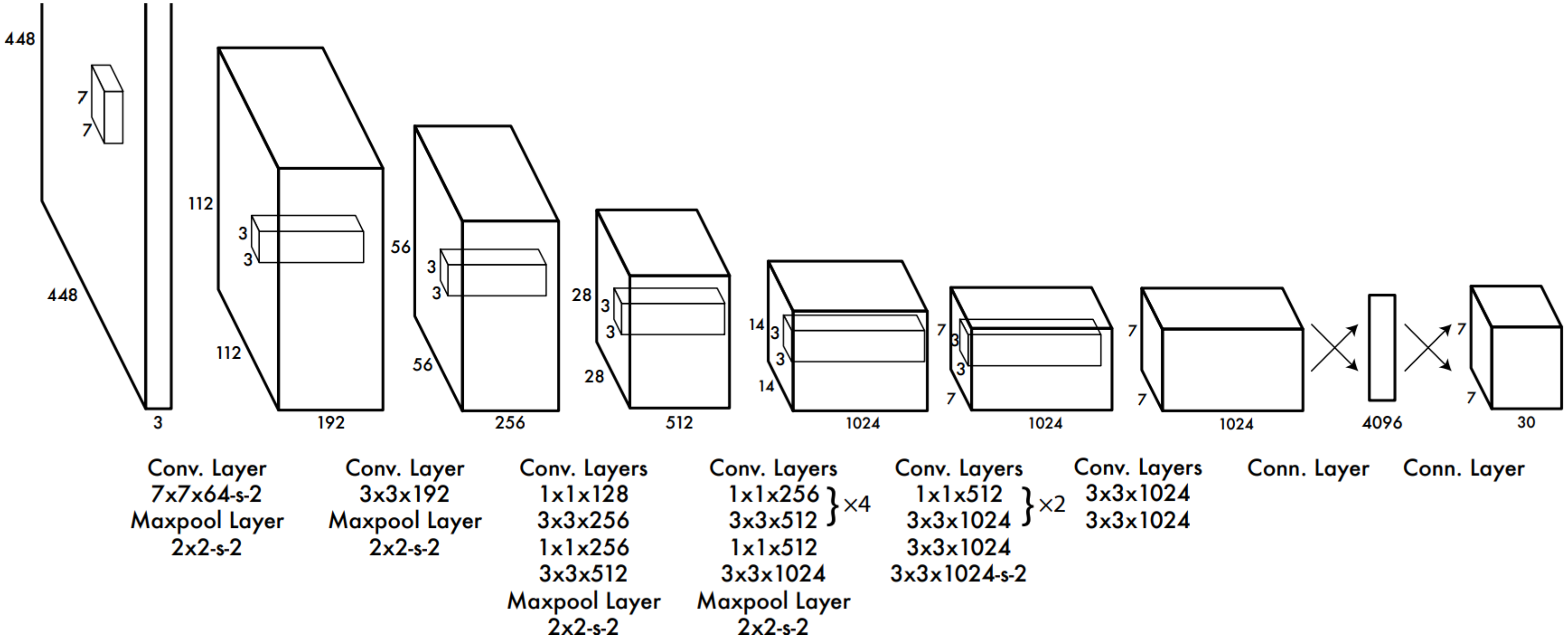
- Detection as a regression problem
- Predicting for each cell:
  - Bounding boxes
  - Confidences
  - Class probabilities
- S x S x (B * 5 + C) tensor



Network Design

- Inspired by the GoogLeNet [2]
- 24 convolutional, 2 FC

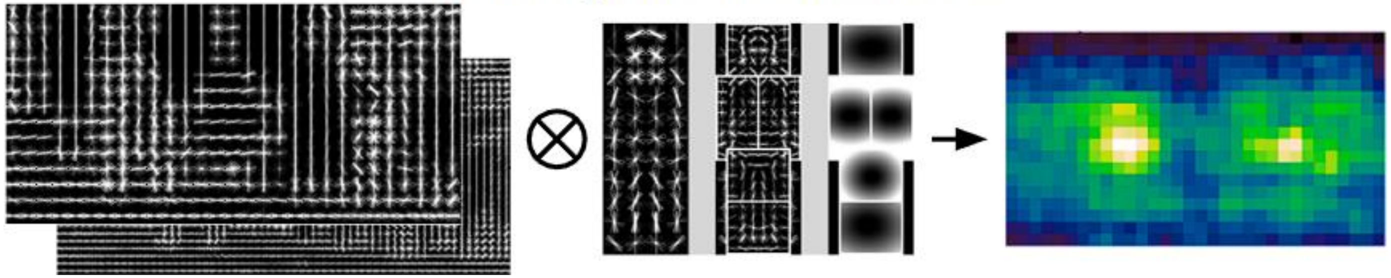
- ~~Inception modules~~
- 1x1 reduction + 3x3 convolutional



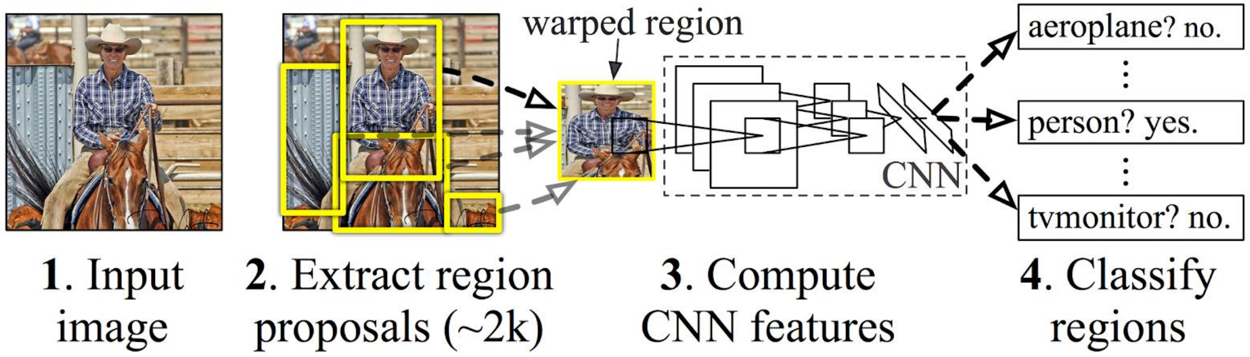
Comparison to Other Detection Systems

Sliding window
DPM
R-CNN
All train region-based classifiers to perform detection

DPM: ***Deformable Part Models***



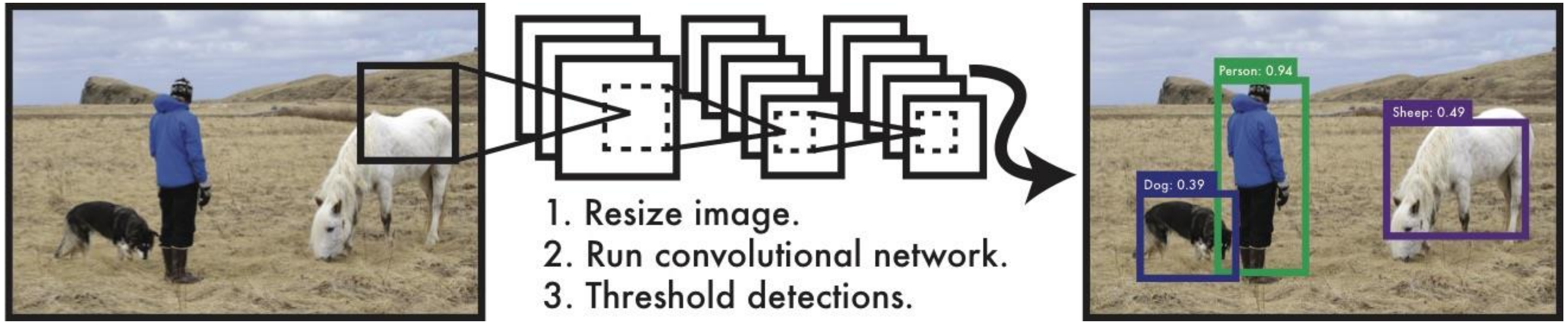
R-CNN: ***Regions with CNN features***



Comparison to Other Detection Systems

With YOLO, you only look once at an image to perform detection

YOLO: ***You Only Look Once***



1. Resize image.
2. Run convolutional network.
3. Threshold detections.

Experiments

- Test bench specs:
  - NVIDIA GeForce Titan X
- No batch processing
- Base network: 45 FPS
- Fast network: >150 FPS (<25 ms latency)

Datasets

- PASCAL VOC 2007 Challenge [4]
- 20 classes:
  - Person
  - Animal: bird, cat, cow, dog, horse, sheep
  - Vehicle: aeroplane, bicycle, boat, bus, car, motorbike, train
  - Indoor: bottle, chair, dining table, potted plant, sofa, tv/monitor
- 9963 images containing 24640 annotated objects

PASCAL2
Pattern Analysis, Statistical Modelling and
Computational Learning

Datasets

- PASCAL VOC 2012 Challenge
- 20 classes
- 11530 images
- 27450 annotated objects

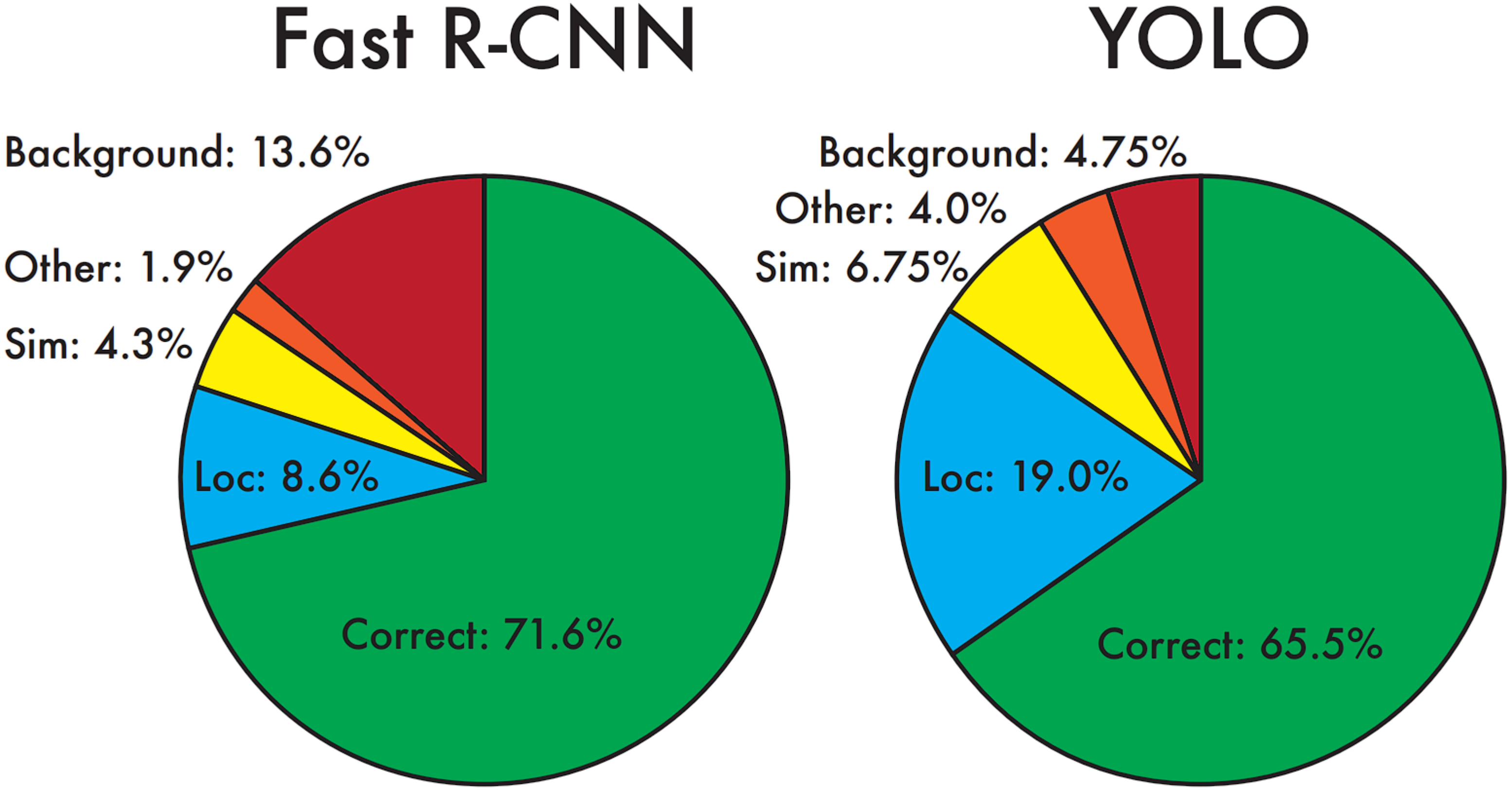
PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning

PASCAL VOC 2007

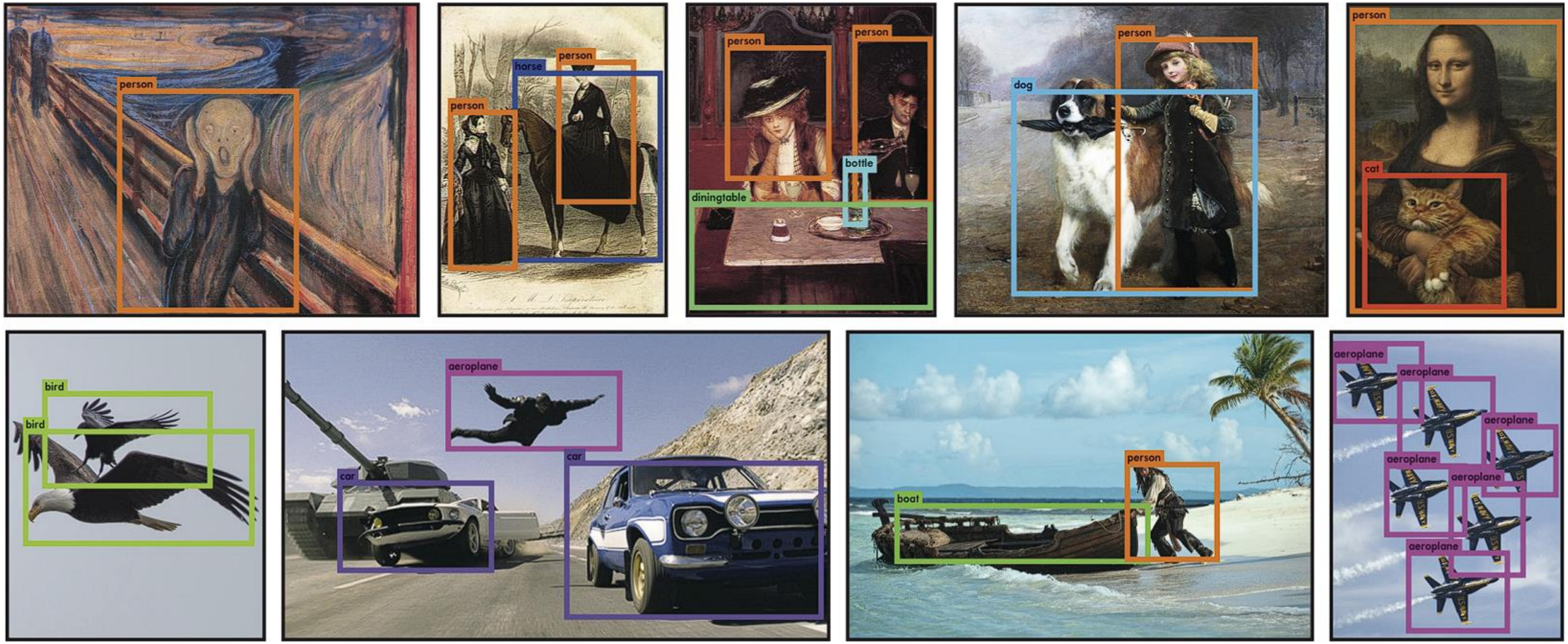
Real-Time Detectors	Train	mAP	FPS
100Hz DPM	2007	16.0	100
30Hz DPM	2007	26.1	30
Fast YOLO	2007+2012	52.7	**155**
YOLO	2007+2012	**63.4**	45
Less Than Real-Time			
Fastest DPM	2007	30.4	15
R-CNN Minus R	2007	53.5	6
Fast R-CNN	2007+2012	70.0	0.5
Faster R-CNN VGG-16	2007+2012	73.2	7
Faster R-CNN ZF	2007+2012	62.1	18
YOLO VGG-16	2007+2012	66.4	21

PASCAL VOC 2007 Error Analysis

- Correct:
  - Correct class
  - IOU > .5
- Localization:
  - Correct class
  - .1 < IOU < .5
- Similar:
  - Class is similar
  - IOU > .1
- Other:
  - Class is wrong
  - IOU > .1
- Background:
  - For any object
  - IOU < .1



Real-Time Detection in the Wild



Object Detection

starring

YOLOv3

Conclusion

- Simple to construct
- Can be trained directly on full images
- Unlike classifier-based approaches, YOLO is trained on a loss function that directly corresponds to detection performance and the entire model is trained jointly

References

[1] J. Redmon et al., "You Only Look Once: Unified, Real-Time Object Detection«, arXiv.org, 2016. [Online]. Available: https://arxiv.org/abs/1506.02640.

[2] C. Szegedy et al., "Going Deeper with Convolutions", arXiv.org, 2014. [Online]. Available: https://arxiv.org/abs/1409.4842.

[3] P. F. Felzenszwalb et al., "Object detection with discriminatively trained part based models", IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(9):1627–1645, 2010.

[4] M. Everingham et al., "International Journal of Computer Vision", 88(2):303-338, 2010.