Person/Object Detection on 360° Images

Restaurant R&D Project

Soumya Chatterjee

Note: Images have been removed due to NDA

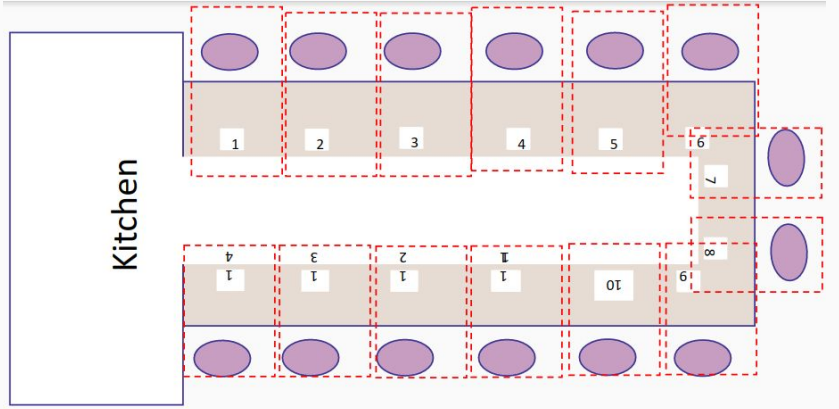
Restaurant Project

About the Project

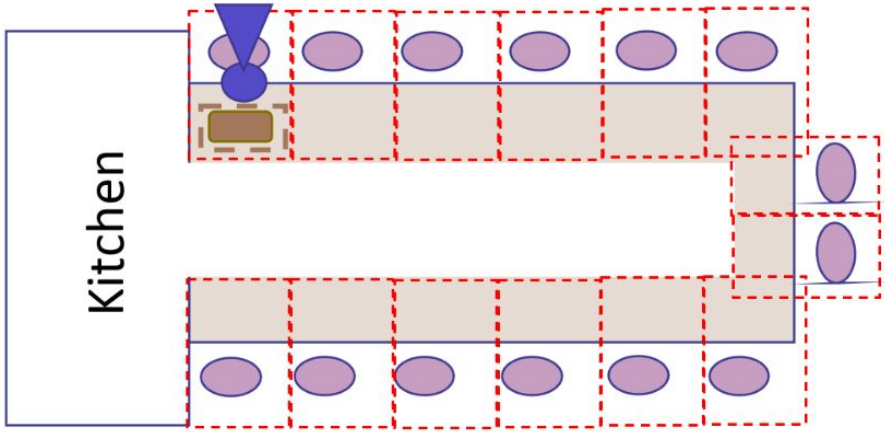
Goals

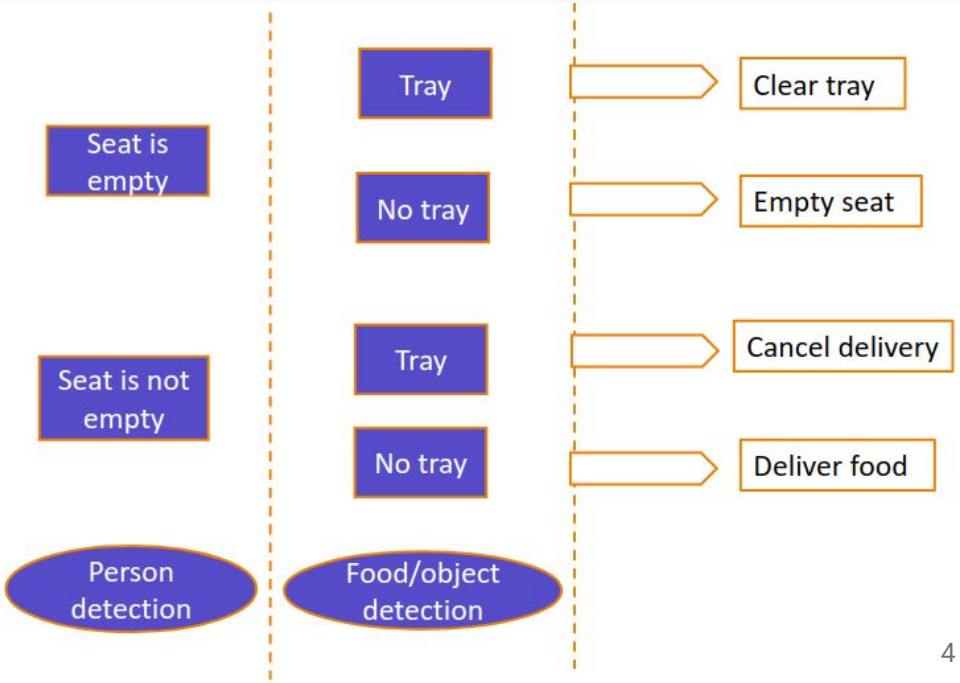
- Finding time elapsed from seating to delivering order
- Detect food delivery to customer and automatically cancel dispatch
- When customer leaves, notification to clear the tray

Fisheye Projection



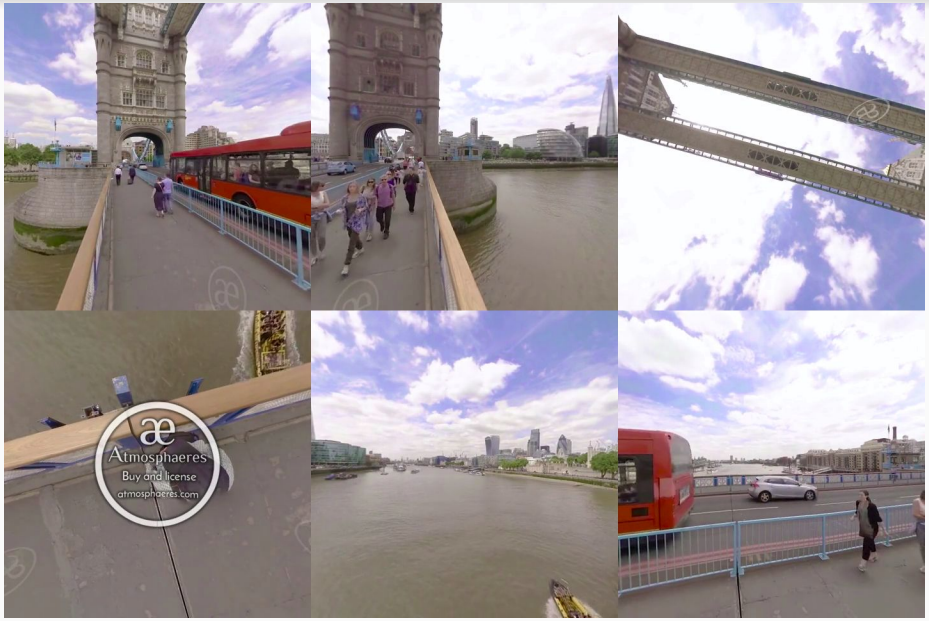
Proposed Method





Different Projections of 360° camera

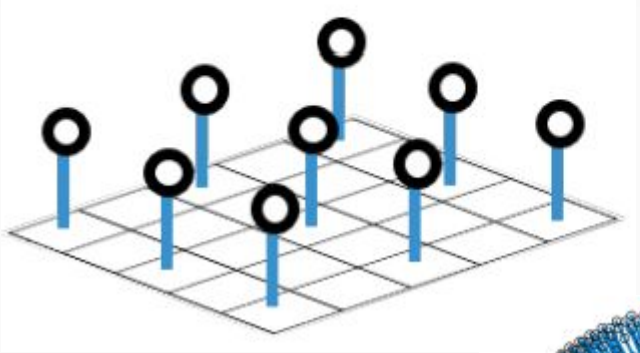
Equirectangular Projection



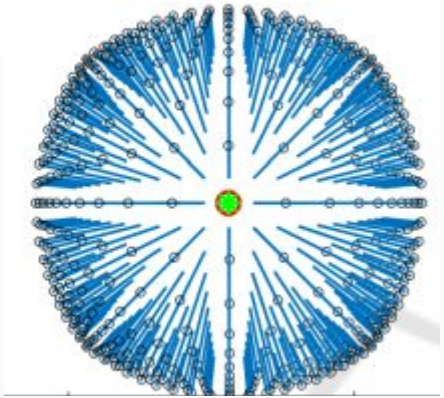
Cube Projection

Challenges with 360° camera

- Objects look very different and distorted in Fisheye projections
- Objects at different locations different orientations in the images
- Existing models do not focus 360° images



Arrangement on Grid



360° Fisheye Projection

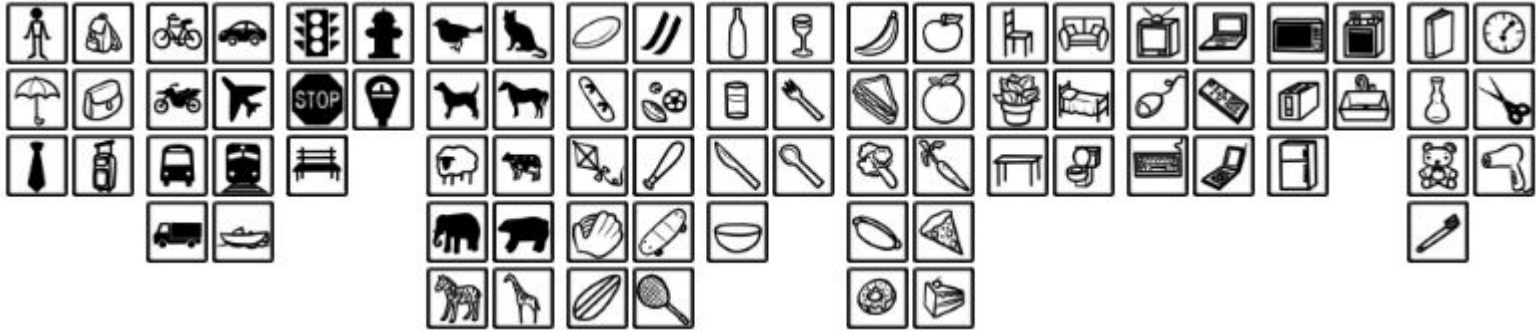
People Detection in Fish-eye Top-views [Demirkus et al.]

COCO Dataset

Common Objects in Context

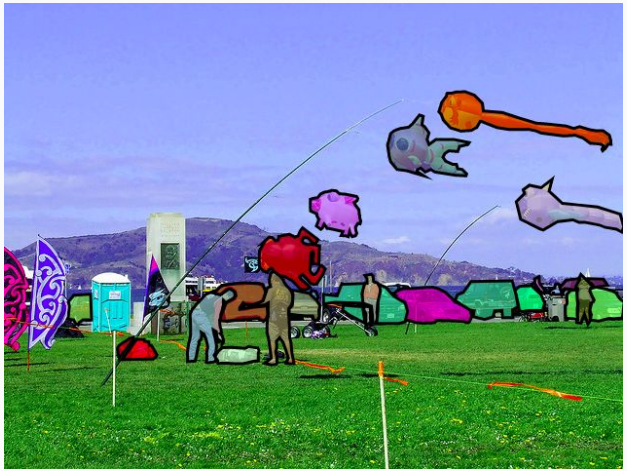
- COCO is a large-scale object detection, segmentation, and captioning dataset
- 200K labeled images
- 80 object categories
- 250000 instances people with keypoints

Supercategories -
outdoor food indoor appliance sports person animal vehicle furniture accessory electronic kitchen



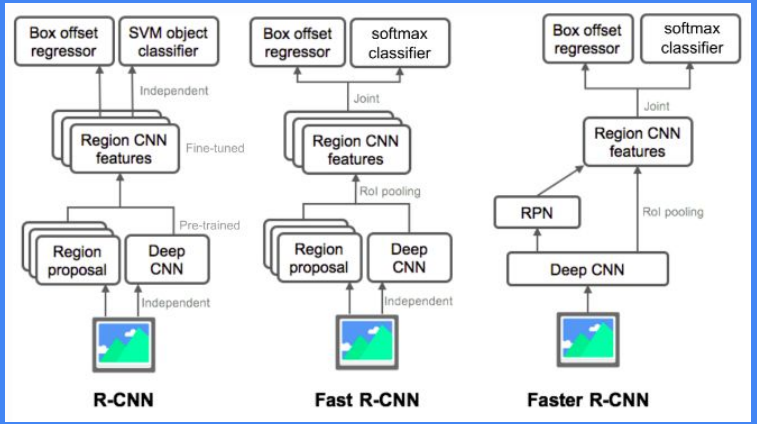
Example Images

Object Detection



Instance Segmentation

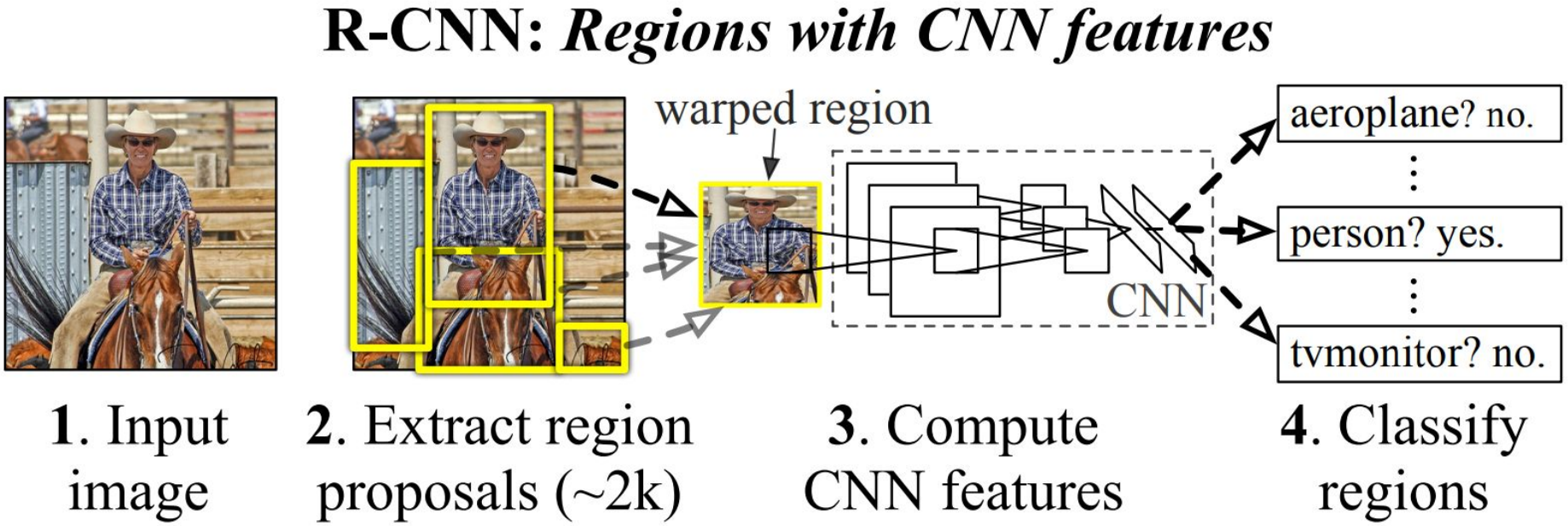
Keypoint Detection



R-CNN Family of Detectors

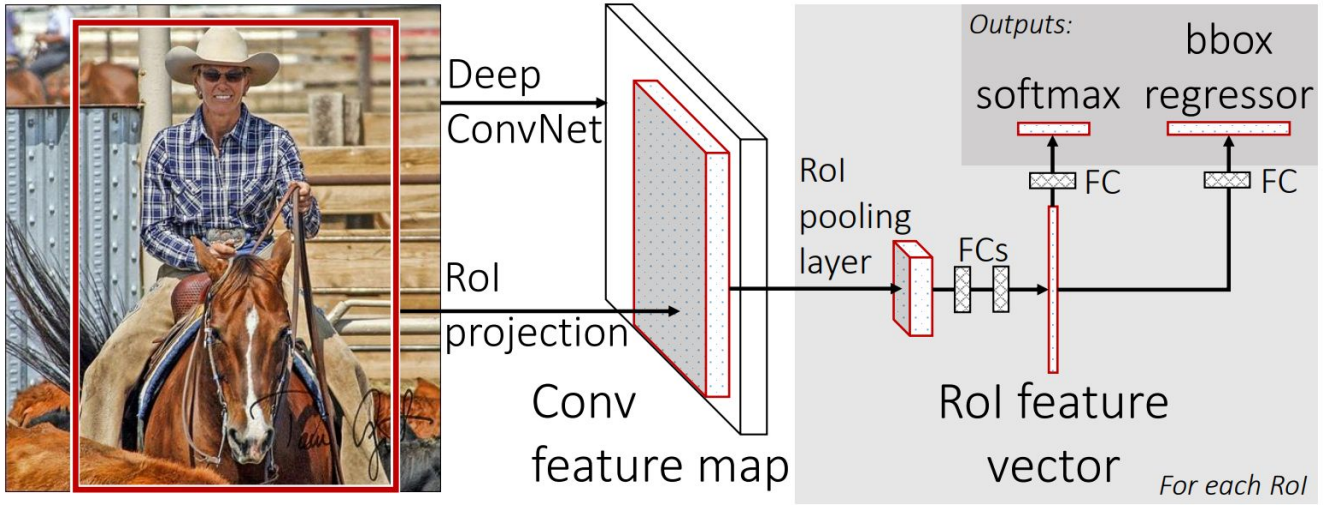
Object Detection for Dummies Part 3: R-CNN Family [Weng, 2017]

R-CNN



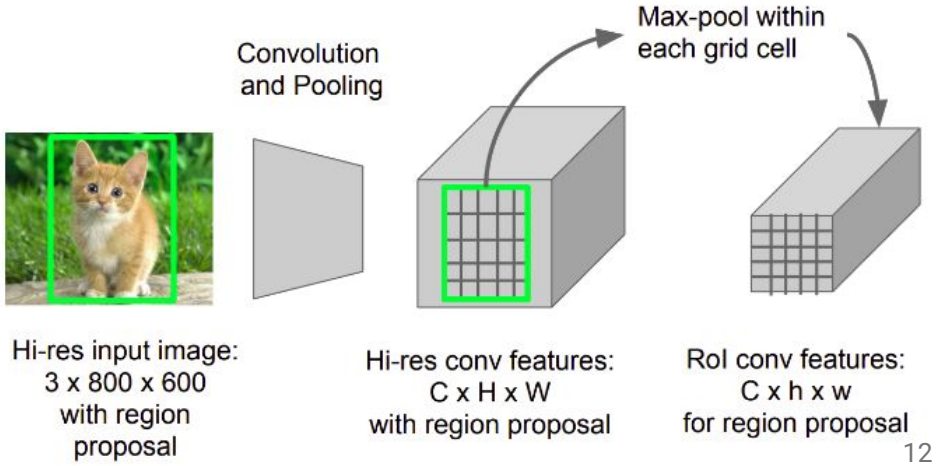
Rich feature hierarchies for accurate object detection and semantic segmentation [Girshick et al., 2013]

Fast R-CNN



Bottleneck in R-CNN -
Generating the CNN feature vector for every image region

Fast RCNN has one CNN forward pass over the entire image and the region proposals share this feature matrix

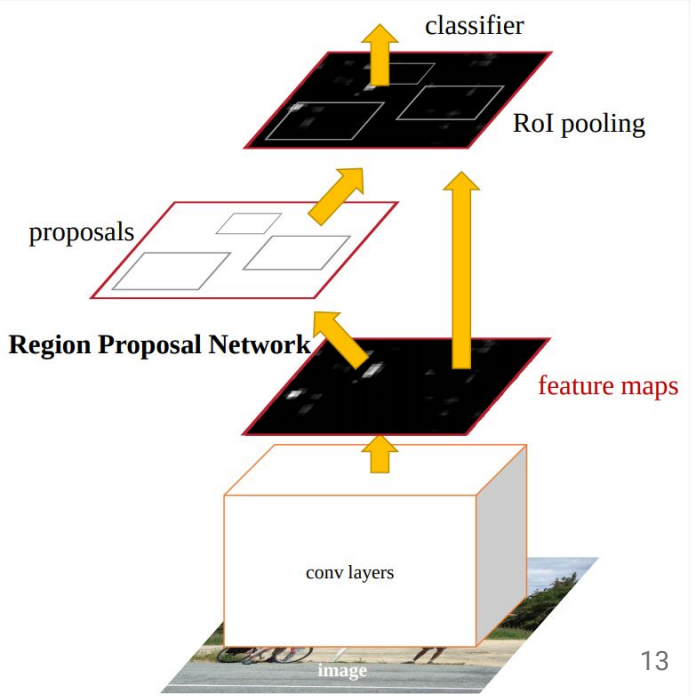


Fast R-CNN [Girshick, 2015]

Faster R-CNN

Bottleneck in Fast R-CNN
Region proposals are generated separately by another model and that is very expensive

Integrate the region proposal algorithm into the CNN model

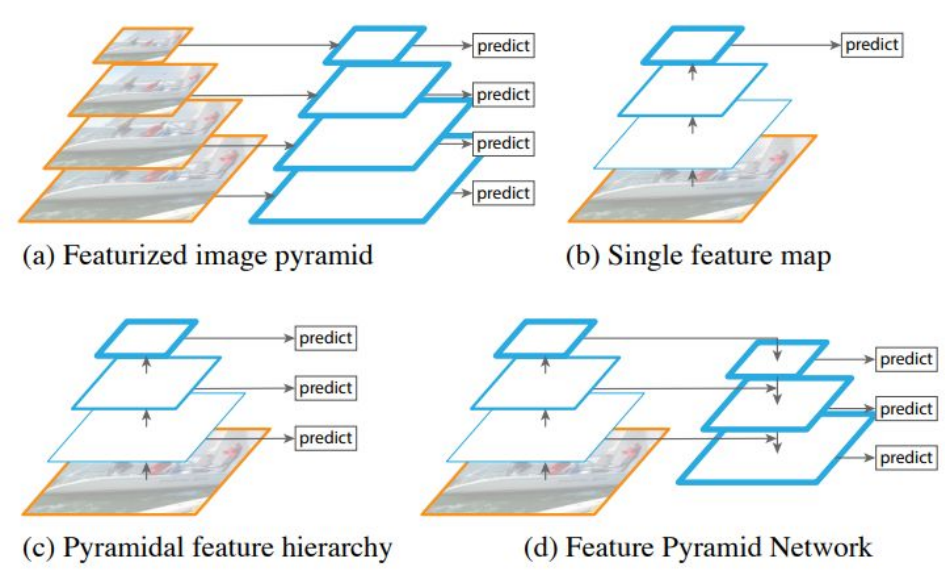


Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks [Ren et al, 2015]

Detectron2
Resnet50-FPN Faster RCNN

Feature Pyramid Network - FPN

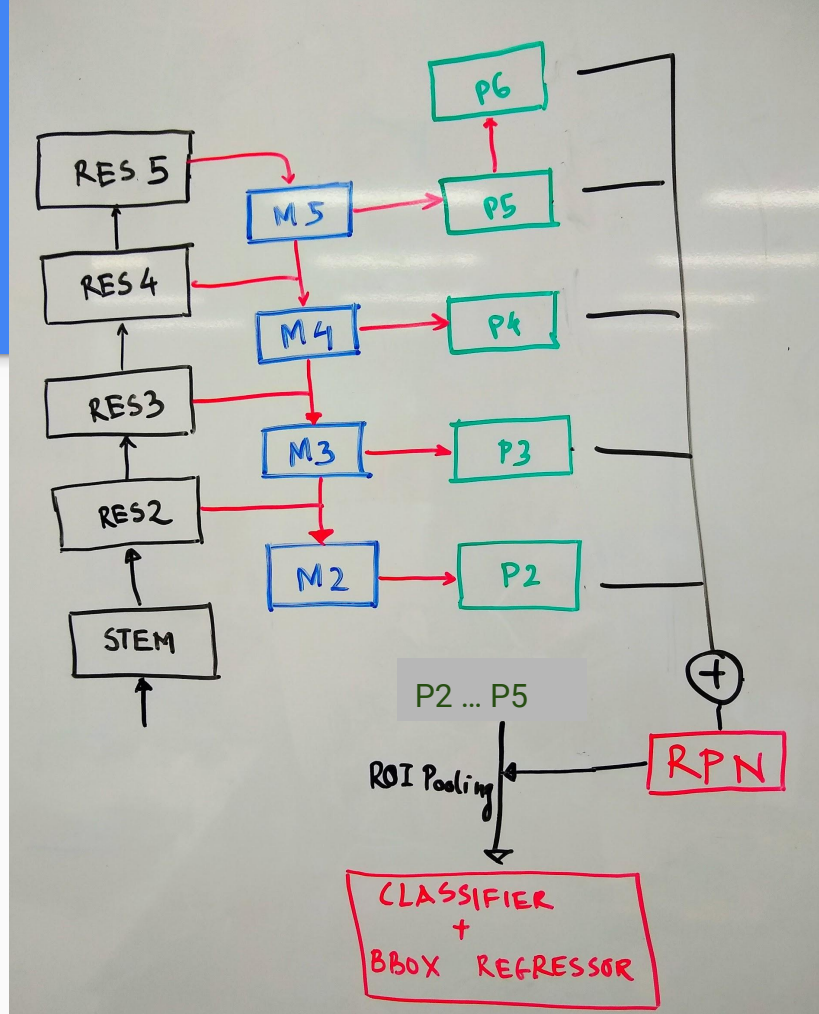
- Feature pyramids were recognition systems for detecting objects at different scales
- Deep learning object detectors have avoided pyramid representations because they are computation and memory intensive
- The paper uses multi-scale, pyramidal hierarchy of deep convolutional networks to construct feature pyramids with marginal extra cost



Feature Pyramid Networks for Object Detection [Lin et al., 2016]

Complete Architecture

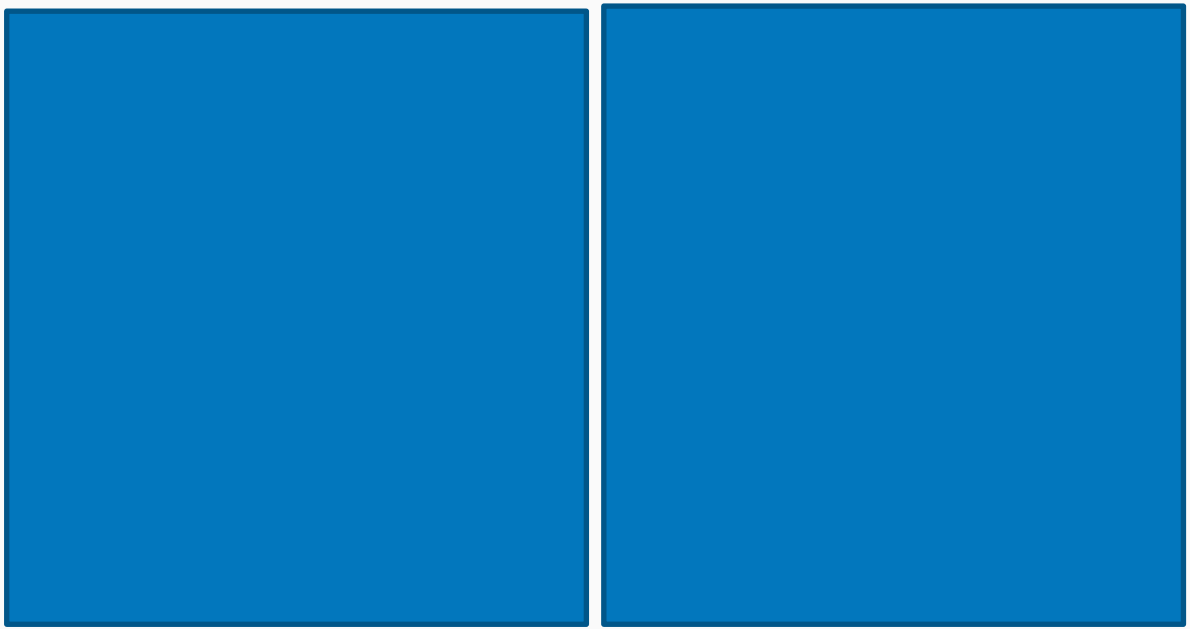
- Resnet 50 Backbone (in black)
- FPN on ['res2', 'res3', 'res4', 'res5']
- RPN uses ['p2', 'p3', 'p4', 'p5', 'p6']
- RoI Heads use ['p2', 'p3', 'p4', 'p5']
- Anchor Boxes
  - Aspect Ratios: [[0.5, 1.0, 2.0]]
  - Sizes: [[32], [64], [128], [256], [512]]



Person Detection

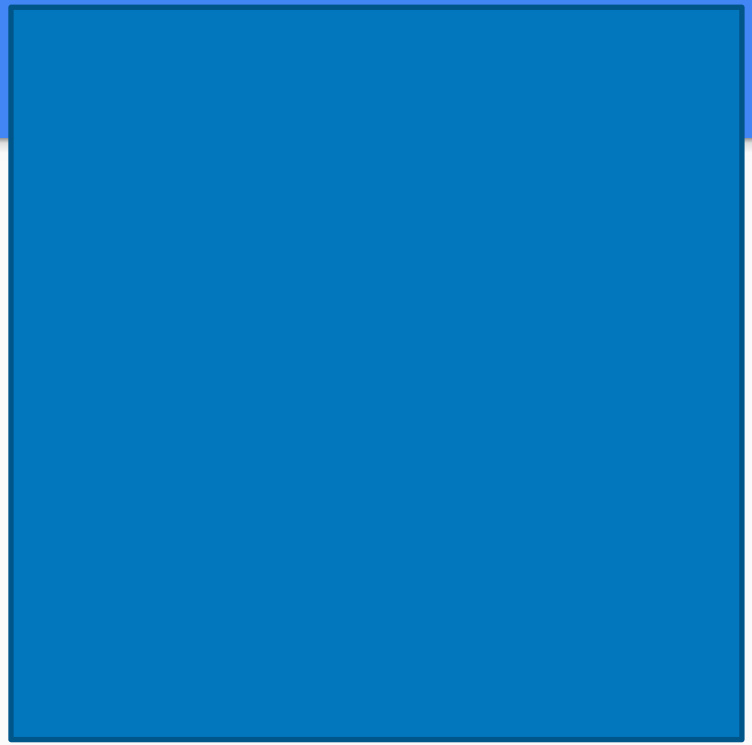
Pretrained Detectron

Results are good only in the upper parts of the image where the 360 images resemble normal images of people in COCO



Finetuning with Augmentation

- Finetuning Dataset:
  - Person images in COCO
- Data Aug:
  - Uniform(-180, 180) rotation
- Prediction Time: 0.05 sec/image
  - Max. achievable FPS ~ 20
- Most people detected but low confidence for people in non-standard orientations



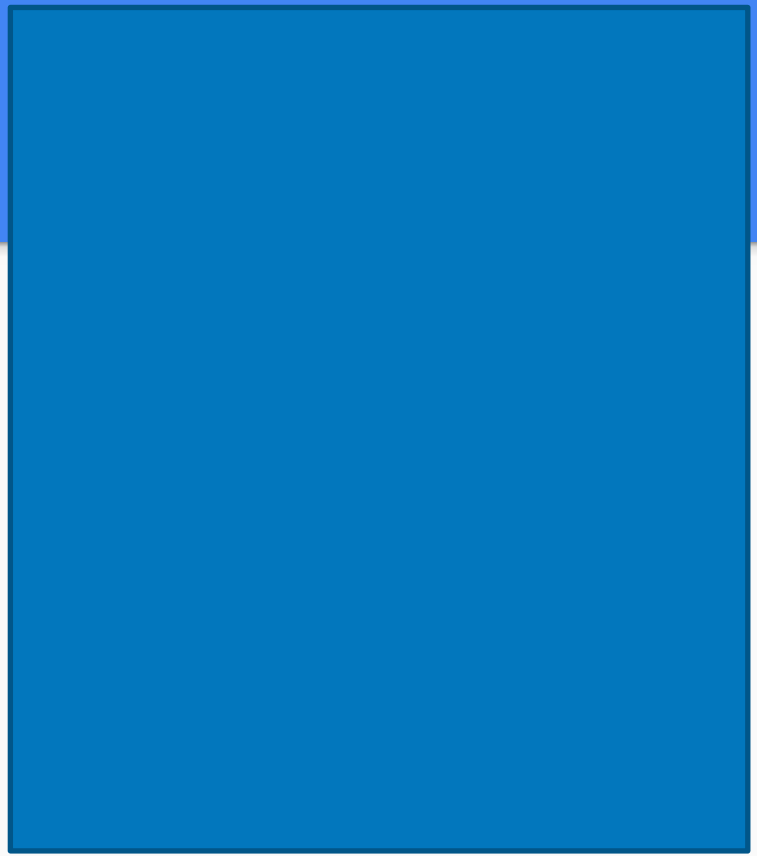
mAP of Person class

	Pretrained	Finetuned
No random rotation during validation	44.29	53.20
Random rotation during validation	15.886 (± 0.31749)	36.68 (± 0.61305)

Head Detection

Detectron Finetuned on Head dataset

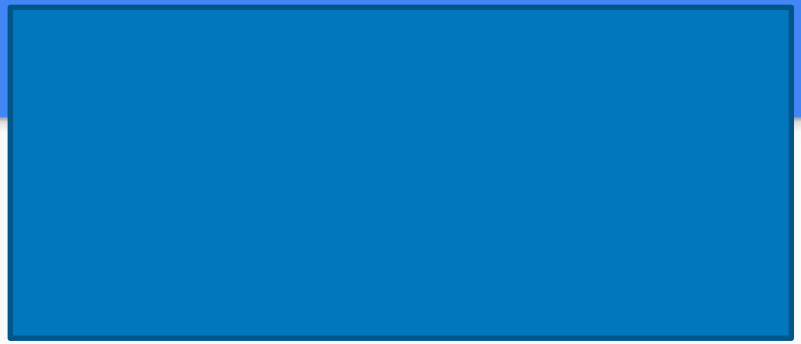
- Motivation:
  - When viewpoint is not proper to view the whole body or it is too crowded, the body annotation can lead to unexpected overlapping bounding boxes
  - Can be avoided by using head annotation
- Observations:
  - Almost all heads detected with high confidence
  - False positives - bowls detected as heads



Object Detection

Detectron Finetuned on Relevant Objects

- Dataset: COCO
  - [bottle, cup, fork, spoon, knife, bowl, cell phone, handbag]
- Data Aug: Uniform(-15, 15) rotation
- Observations
  - Detection results are poor
  - Some bowls get detected but as cups
- Possible reason for failure
  - Absence of proper dataset



Multi-label Classification (WIP)

Instead of Person and Object Detection, multi-label classification can be done in each service area

Preliminary Experiment

- COCO Dataset - Apples and Bananas classes
- Simple VGG Net + FC layer
- Accuracy (similar to IoU) ~ 70%
- Per class accuracy ~ 70% for both

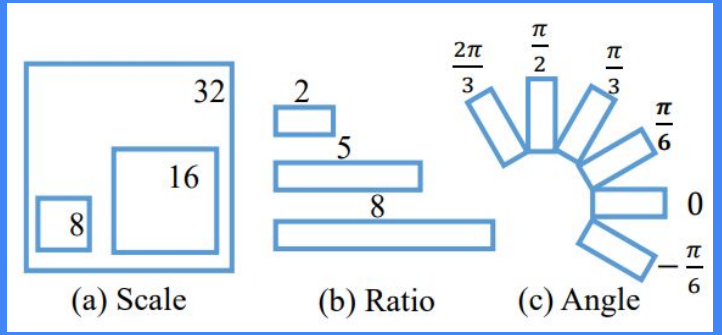
Need to find suitable dataset, losses and evaluation metrics

tensor([0., 1.]) tensor([0., 1.])

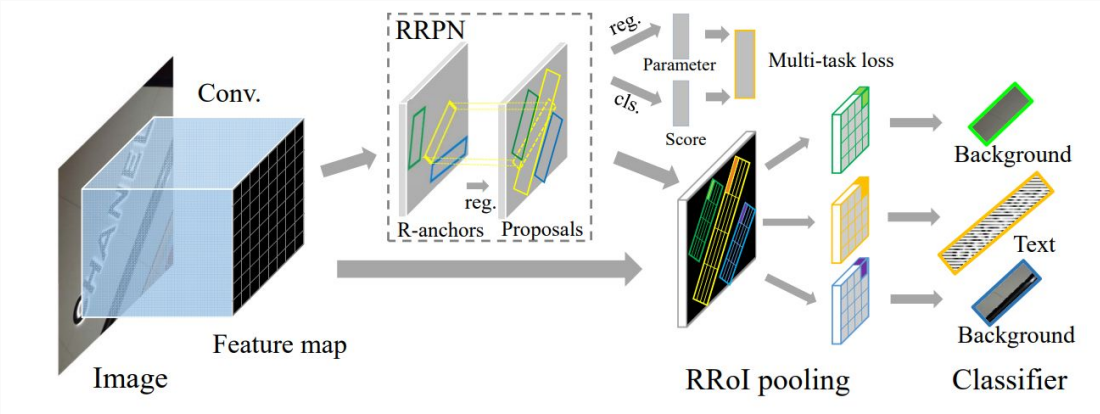
tensor([1., 0.]) tensor([1., 0.])

tensor([1., 1.]) tensor([0., 1.])

Future Work



- Compare with other pretrained models of Detectron
- Use a Rotated RPN
- Improve augmentation by using segmentation maps for tighter bounding boxes
- Detection on Fisheye images converted to hemi cube



Arbitrary-Oriented Scene Text Detection via Rotation Proposals [Ma et al., 2017]

Thank You