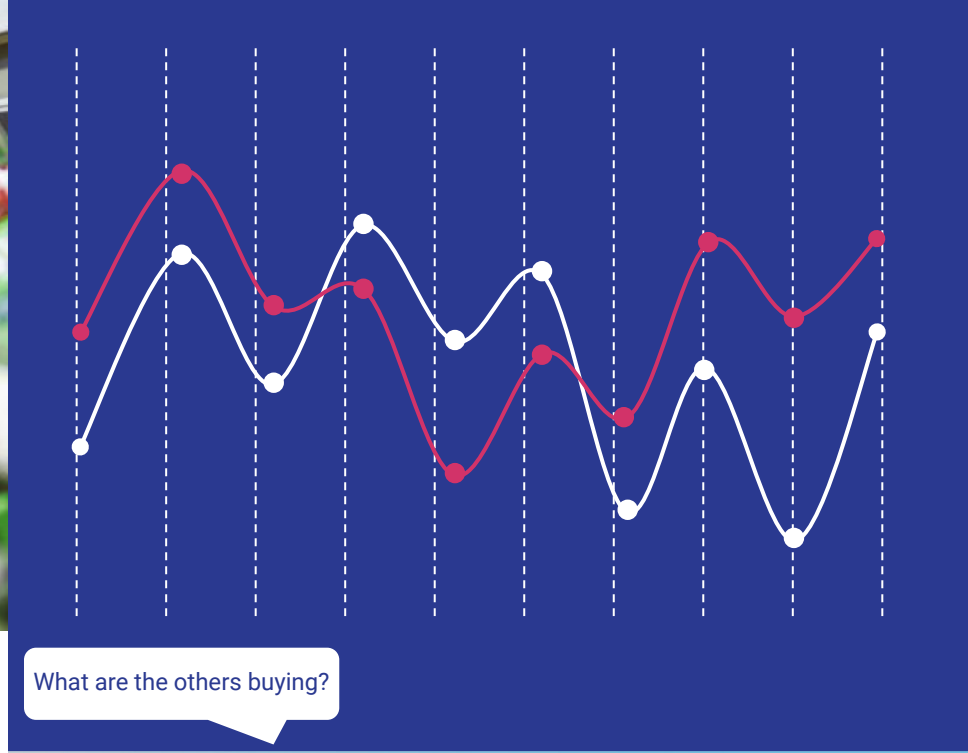
Human-Object Interaction in Retail

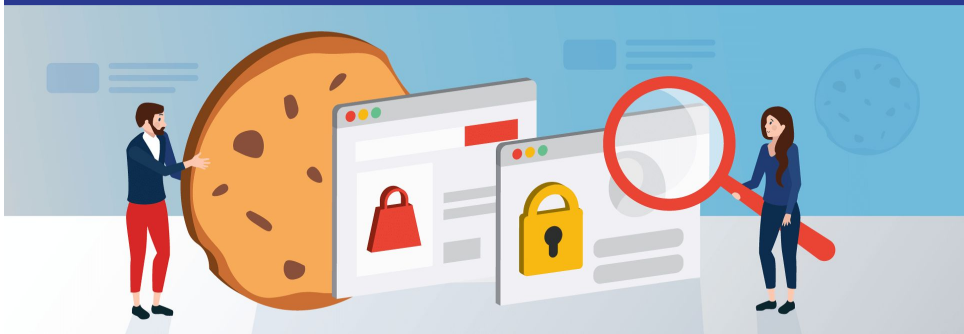
Max Tanski & Lauren Heintz
6.869 Computer Vision Final Project

Purpose

The Rise of Retail Cookies

Capturing human object interaction in grocery stores

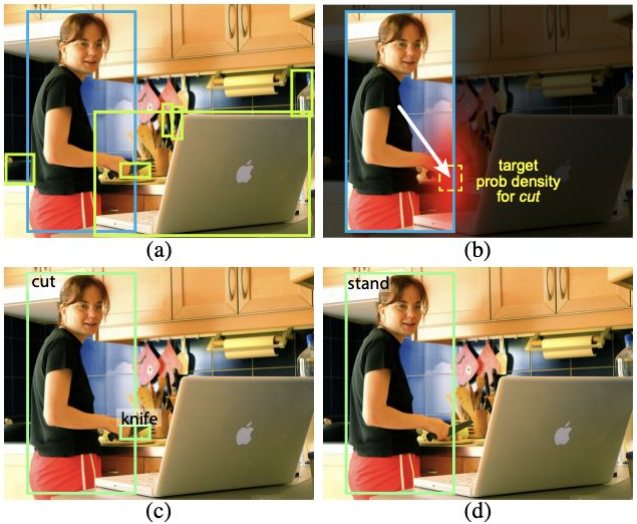




State of the Art Approaches

< human, action, target >

Instance Segmentation & Pose Estimation

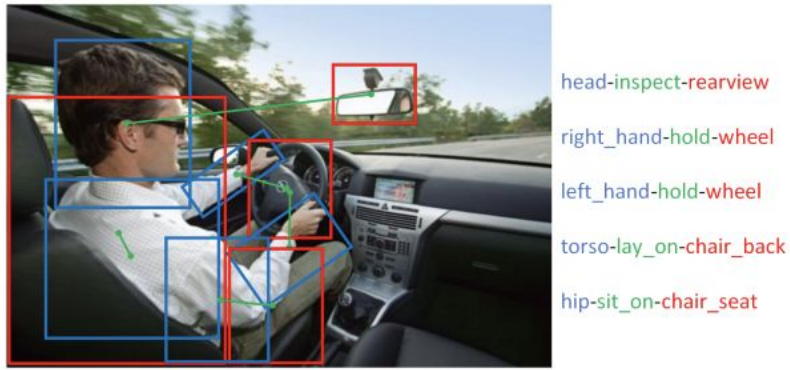


FacebookAI InteractNet

< Person, Cut, Knife >

< human, action, target >

Keypoints & Part State Relationships

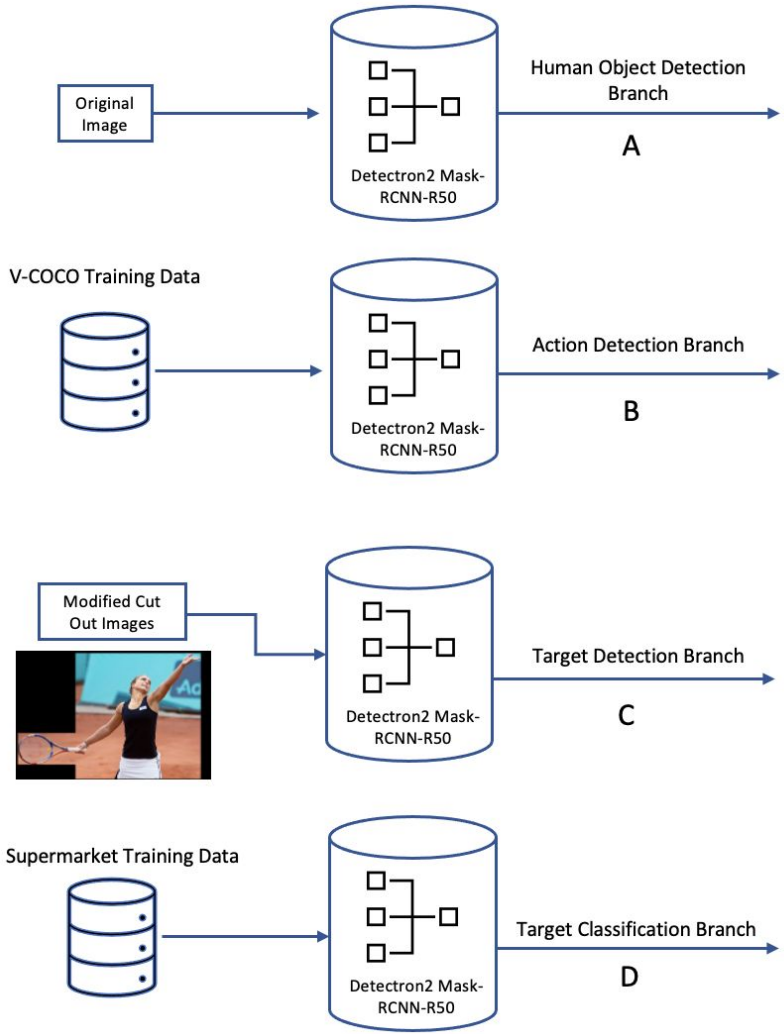


HAKE: Human Activity Knowledge Engine

< Person, Drives, Car >

< human, action, target >

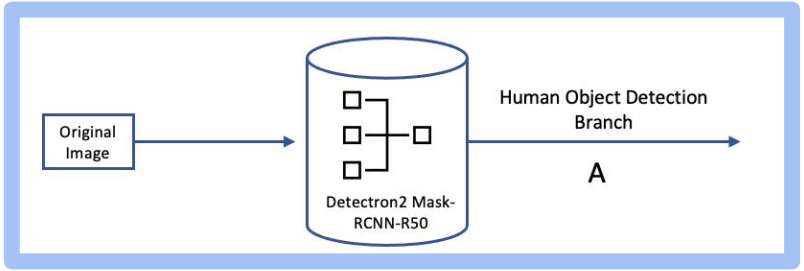
Approach



Our Approach

< human, action, target >

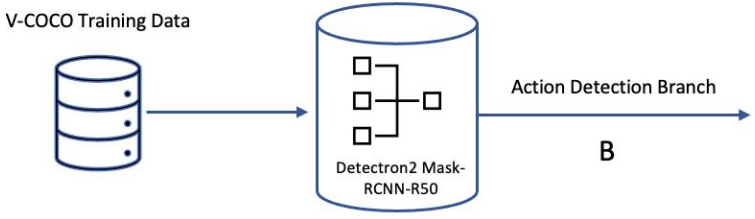
A. Human Object Detection Branch
B. Action Detection Branch
C. Target Detection Branch
D. Target Classification Branch

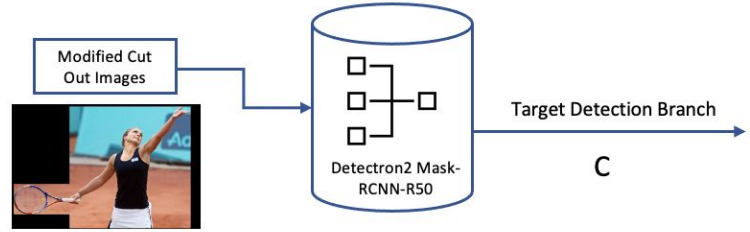


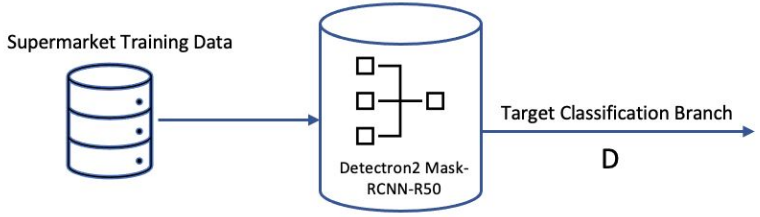
1. Person Label
2. Person Box
3. Person Mask
4. Object Label
5. Object Box
6. Object Mask

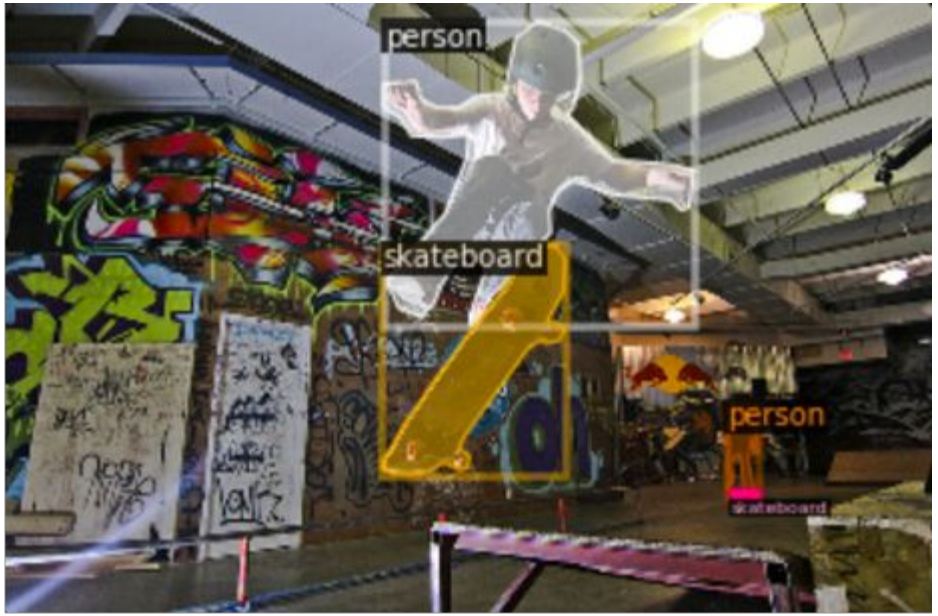
Branch A

< human, ~~action~~, target >





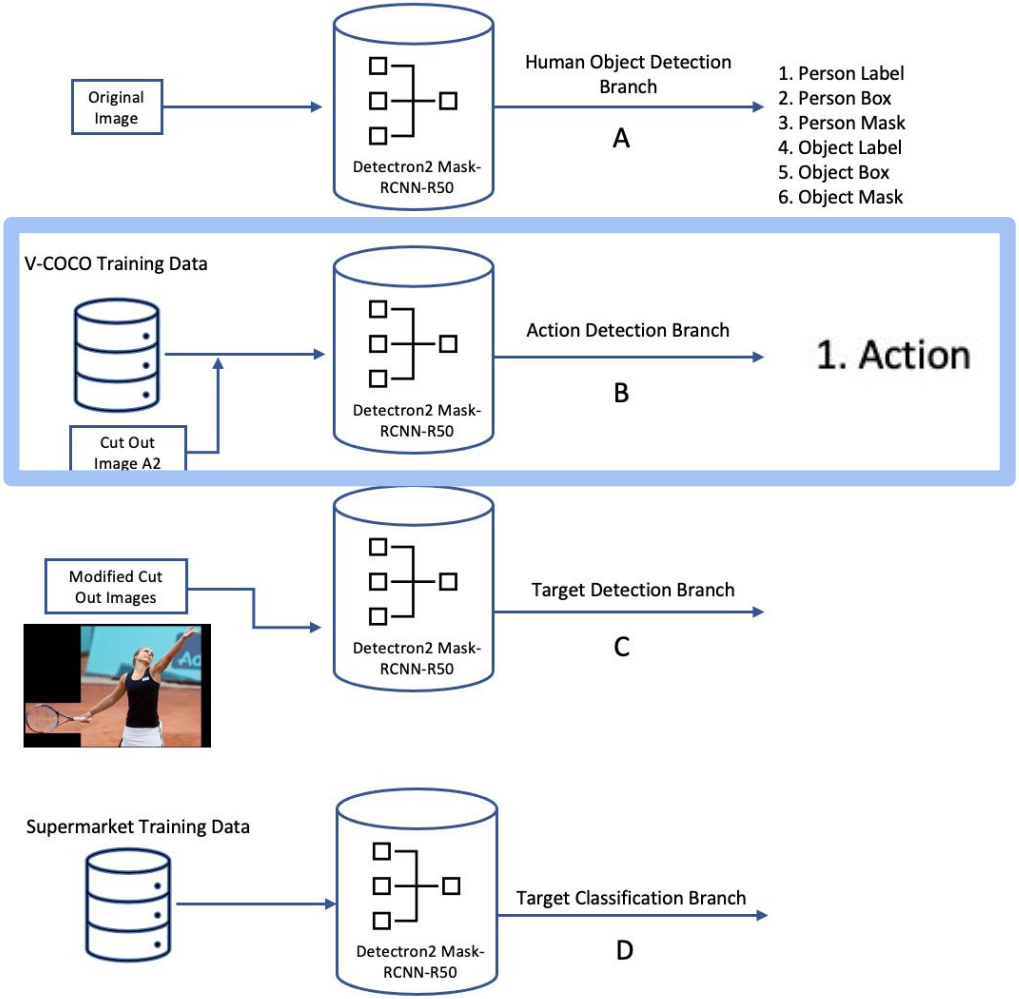




Branch A


person 100%
orange 86%
orange 88%
orange 81%

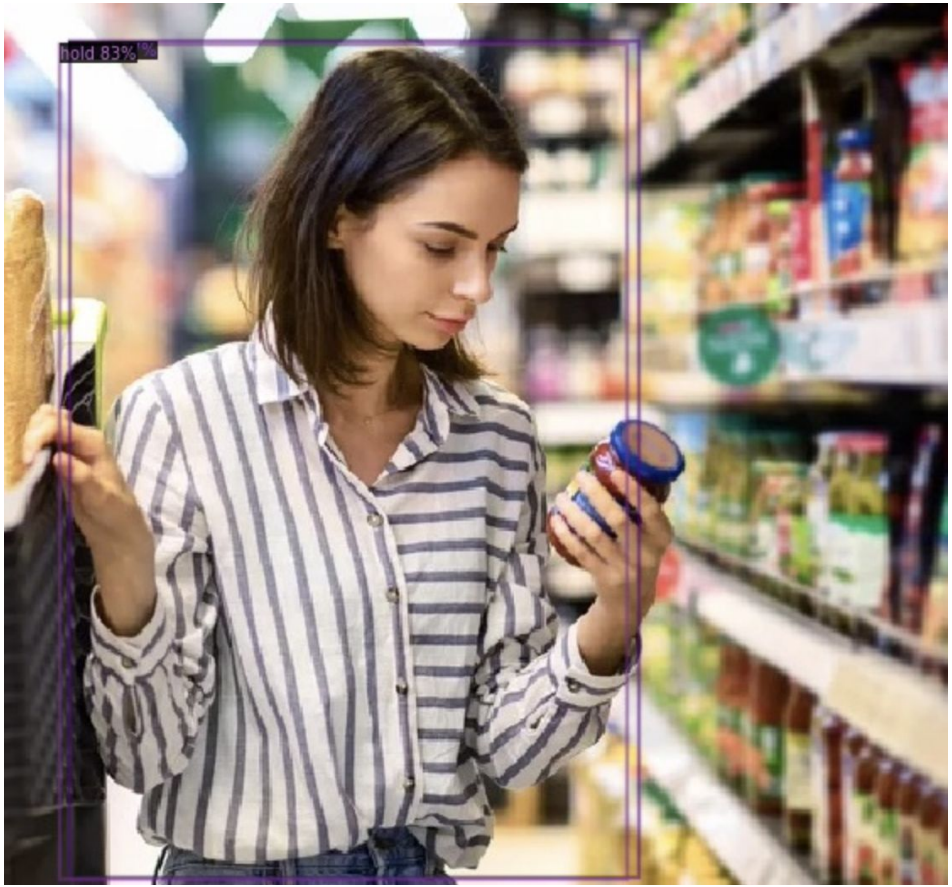

person 100%
person 62%
person 98%

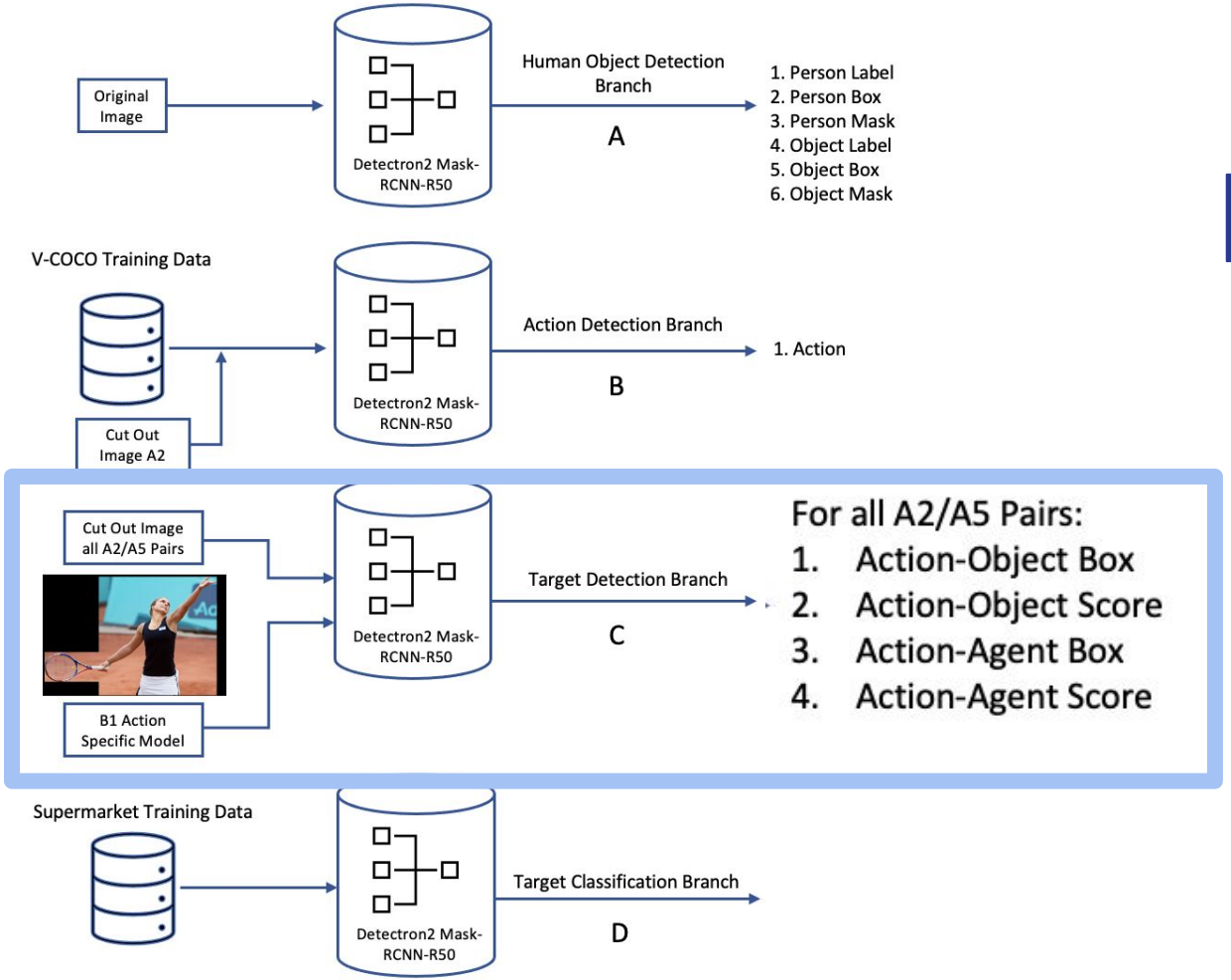


Branch B

< ~~human~~, action, ~~target~~ >

Branch B





Branch C

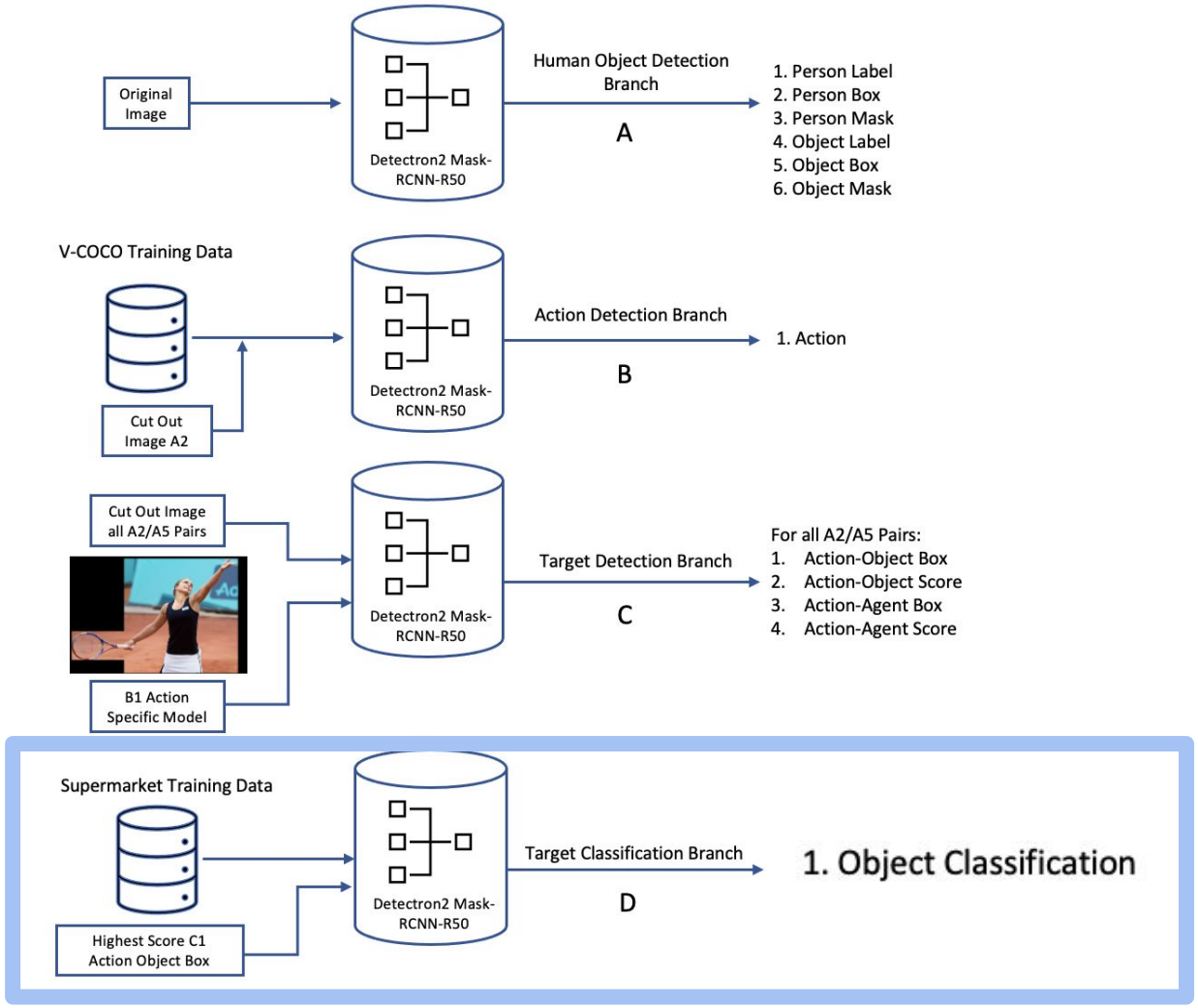
< human, ~~action~~, target >

Branch C


hold_obj 87%
hold_obj 83%
hold_obj 74%


hold_obj 83%
hold_obj 81%
hold_obj 93%
hold_obj 84%
hold_obj 85%


hold_obj 84%
hold_obj 81%



Branch D

< ~~human~~, ~~action~~, target >

Branch D


oranges 27%
oranges 11%
clementine_single 16%


lettuce 18%


oranges 15%
orange_single 41%
orange_single 25%
oranges 11%


lettuce 11%
orange_single 11%12%

Results

Experiment Parameters

Backbones

Types tested:

`mask_rcnn_R_50`

`faster-rcnn_R_50`

`retinanent_R_101`

Epochs

Values tested:

`100`

`300`

`500`

`1000`

`1200`

Learning Rate

Values tested:

`0.01`

`0.002`

`0.001`

Data Sets

Types:

`COCO`

`V-COCO`

`V-COCO Agents`

`V-COCO Cutouts`

`Grocery Data`

*Cross Entropy Loss Solver and Adam's optimizer kept constant

Metrics for Action Detection and Object Detection

Average Precision

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

TP = True positive

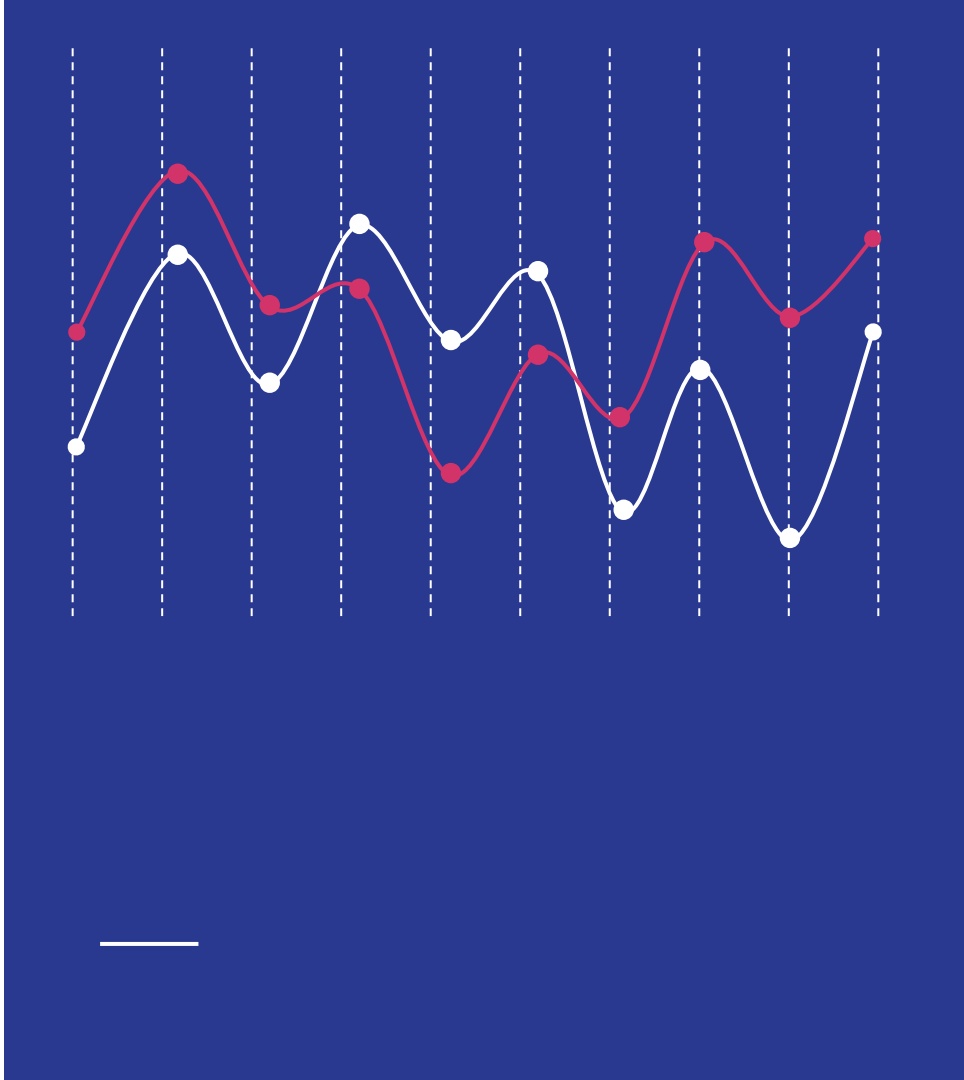
TN = True negative

FP = False positive

FN = False negative

Mean Average Precision

$$mAP = \frac{1}{|classes|} \sum_{c \in classes} \frac{|TP_c|}{|FP_c| + |TP_c|}$$



Branch mAP Best Results

Branch A

Model:

`mask_rcnn_R_50`

Data: COCO

Metrics:

AP = 53.6

mAP = 39.9

Branch B -

Model:

`retinanent_R_101`

Data: V-COCO Agent

Metrics:

AP = 14.973

mAP = 18.309

“Hold” AP = 37.947

Branch C

Model:

`faster-rcnn_R_50`

Data: V-COCO Cutout

Metrics:

AP = 12.776

mAP = 14.092

“Hold_obj” AP = 38.219

Branch D

Model:

`faster-rcnn_R_50`

Data: Groceries

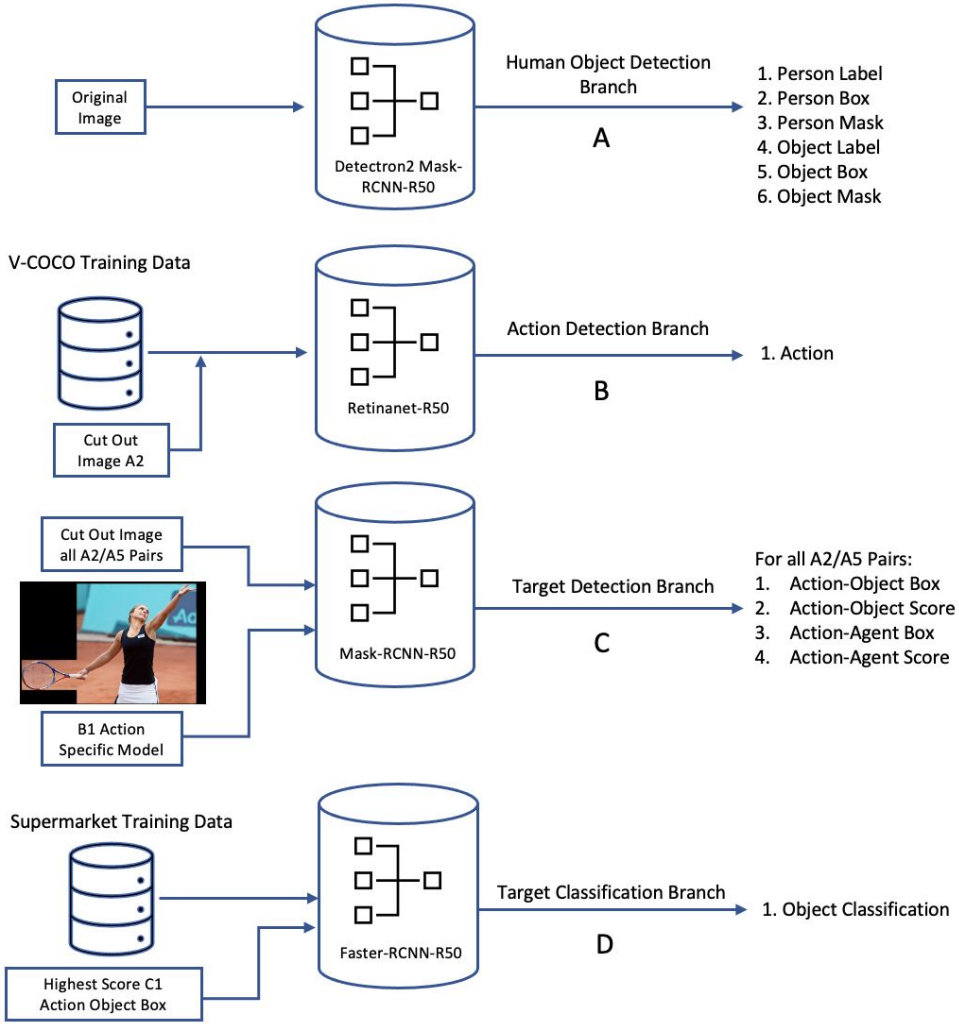
Metrics:

AP = 23.377

mAP = 9.414

*State of the art benchmarks AP = ~48

Example Output of Pipeline



Final Architecture

< human, action, target >

A. Human Object Detection Branch
B. Action Detection Branch
C. Target Detection Branch
D. Target Classification Branch

Thank you! Questions?