Real-Time Computer Vision in Retail

NVIDIA GTC 2019

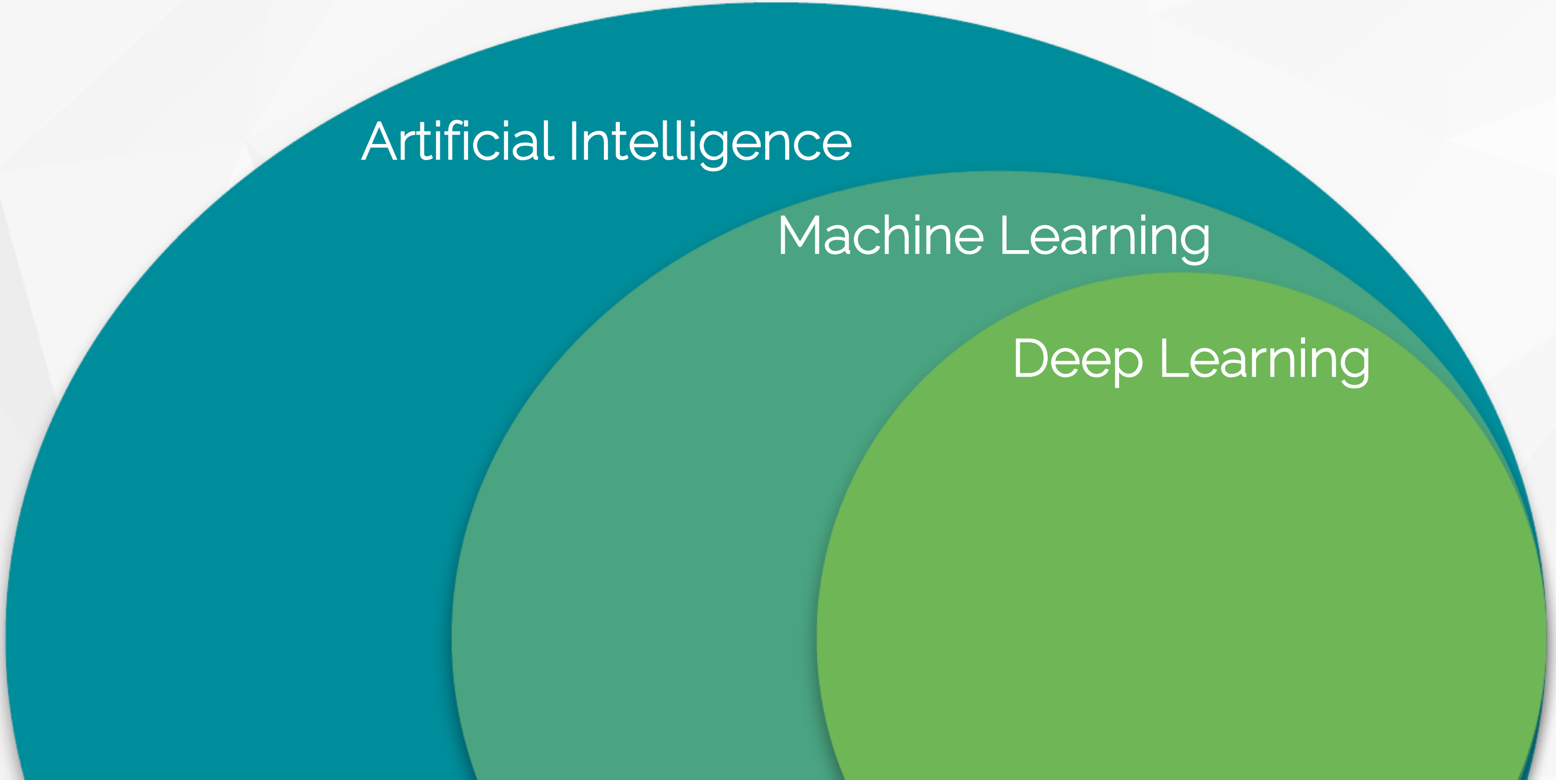
Version: 8 March 2019

Agenda

- AI in Retail
- Real World Challenges
- Technical Obstacles
- Case Study: *Inference at the Shelf*
- Future Areas of Research

AI in Retail

Note



“Just Walk Out”

Source: SounderBruce – Creative Commons Attribution-Share Alike 4.0 International license.

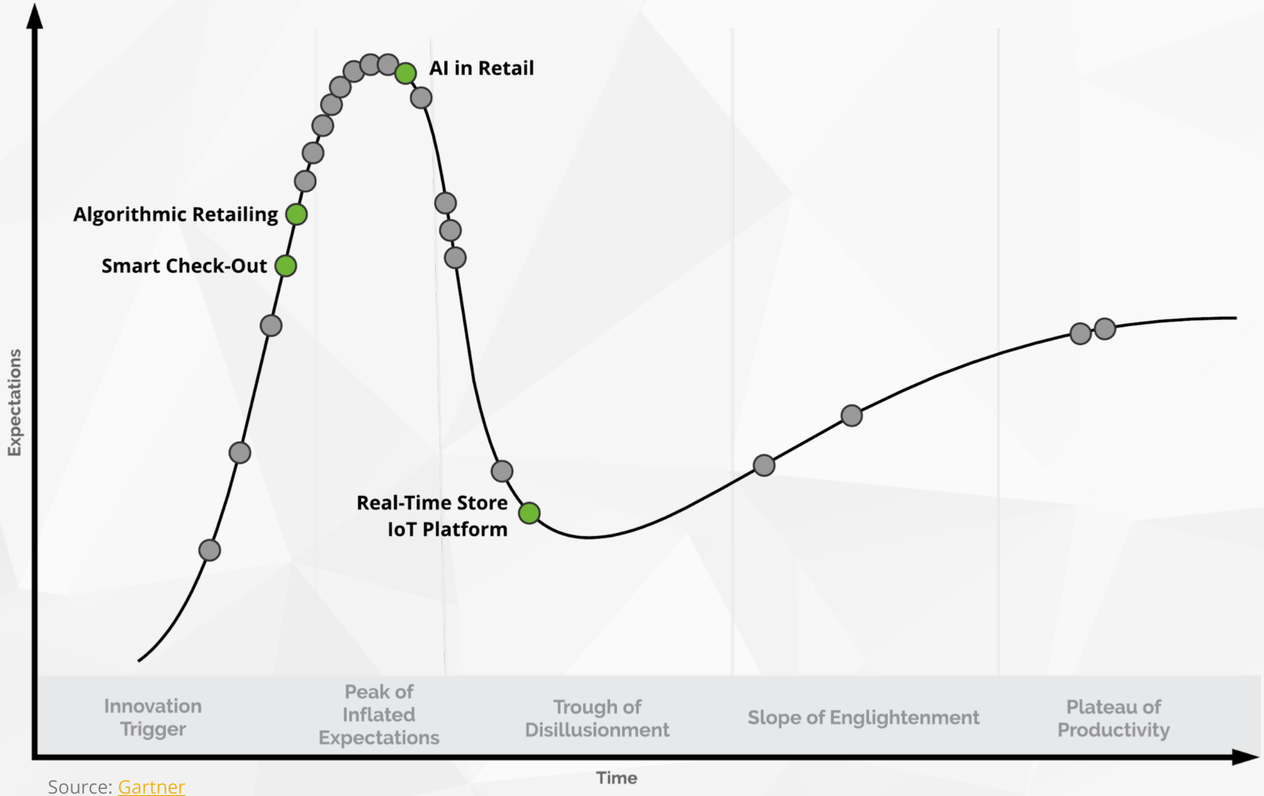
Retail Apocalypse!

Amazon is killing retail…

Successful retailers will…

- feature exclusive products.
- resurrect the art of selling.
- deliver a satisfying experience.
- challenge the fundamental assumptions of commerce.

Retail Technology Hype Cycle



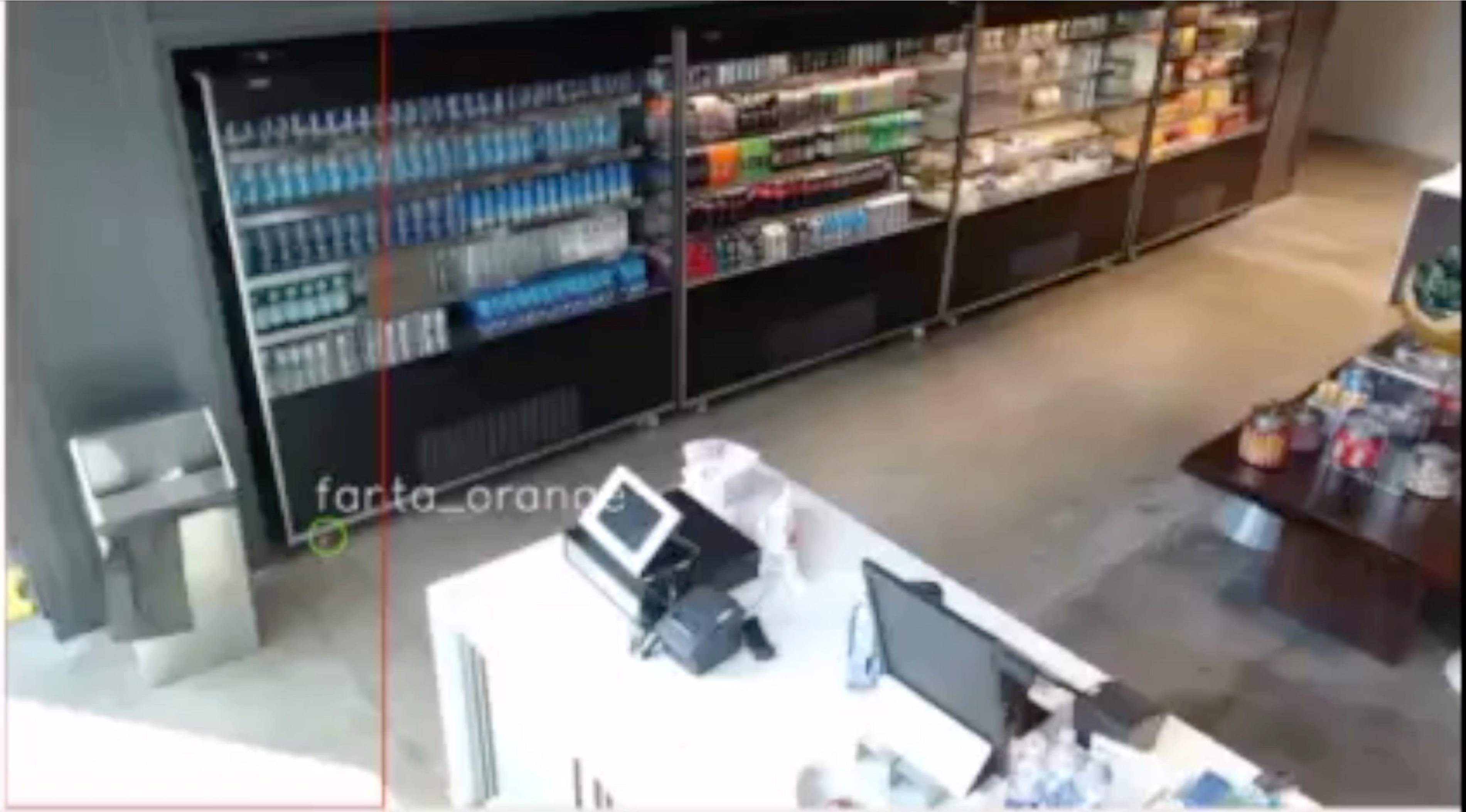
Source: Gartner

Observations at NRF

NRF 2019
RETAIL'S BIG SHOW

- Smart shelves
- People tracking
- Item detection
- Fraud / shrink detection & prevention
- Smart carts
- Age verification

Frictionless Consumer Experience


fanta_orange

Real World Challenges

Business and Operational Challenges

Consumer Experience	Store Redesign	Privacy	RoI
▪ Frictionless, SCO, and assisted checkout ▪ Opt-in vs. Opt-out	▪ Aisles ▪ Power & networking ▪ Minimize occlusion	▪ Always on camera? ▪ Children on camera? ▪ Right to be forgotten?	▪ Cost/Benefit of frictionless ▪ Ways to drive value without increased cost? ▪ Empower over curated?

Technical Challenges

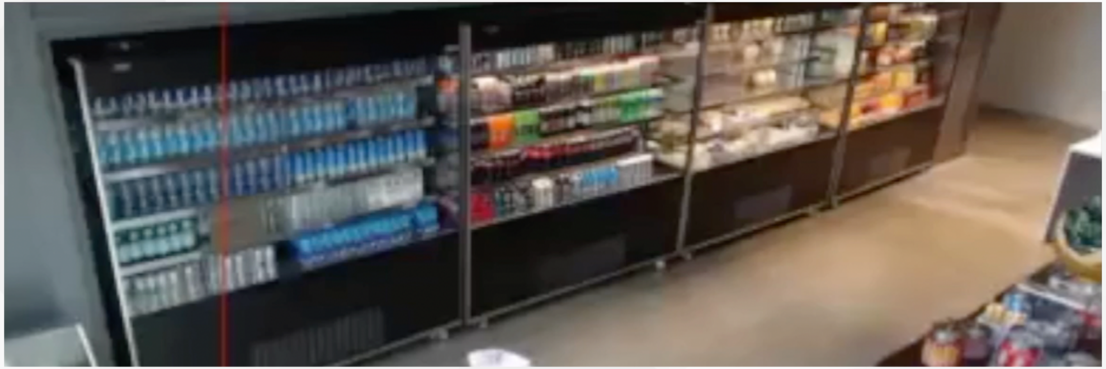
Technical Challenges

- People detection and tracking
  - How can I track people who appear to be very similar? (twins, uniformed, etc.)
  - How do I differentiate between shoppers and employees
  - How do I handle multiple shoppers with a shared cart?
  - Shoppers with children.
- Item detection, recognition, and tracking
  - New items, small items, similar items
  - Carts vs. bags
- Other obstacles
  - Occlusion of people and items
  - Real-time & latency
  - Consequences of false positives, false negatives, etc.

Inference at the Shelf

case study

Not our first rodeo



Problem Statement

One approach to offering a frictionless shopping experience is to recognize items removed from a retail shelf and automatically add them to a shopper's virtual cart in real-time.

Key Requirements

Detection failures result in giving items away for free.

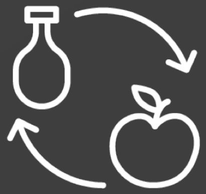
Cart-to-person mismatches result in freebies and erroneous charges.

Recognition failures result in charging for the wrong items.

Sub real-time processing misses add to & remove from cart events.

Key Requirements

Detection failures result in giving items away for free.



Cart-to-person mismatches result in freebies and erroneous charges.

Recognition failures result in charging for the wrong items.

Sub real-time processing misses add to & remove from cart events.

Approach

Use computer vision and deep learning for object detection and classification and NVIDIA GPUs to accelerate inference to achieve real-time performance.

Motivation

A deep neural network trained on <u>thousands</u> of **low resolution** images with a *distribution resembling the validation set* is more likely to have high detection and recognition accuracy as well as perform real-time inference at a high frame rate.

Practical Challenges

Assembling a well-distributed dataset

- How many samples per class are needed?
- In retail, appearance changes frequently
- Annotation cost
- Annotation time
- Manual or automated data acquisition -> labeling pipeline?

Practical Challenges

Selecting a neural network architecture for this use case

- Complex discussion beyond the scope of this talk

Practical Challenges

Performing inference in real-time

- Experiment with smaller image resolutions to improve FPS
- Test different GPUs
- Edge vs. centralized processing

Practical Challenges

Achieving accuracy suitable for the use case

- Connects back to the key requirements we discussed previously
- Missing or incorrectly classifying items has serious implications in retail

Practical Challenges

Cameras

- Sensor types
- Lenses
- Mounting height
- Field of view
- Pixels per inch (PPI)

Experimental Variables

Experimental Variables

- Which combination is best and why?
- Experiment evaluates varying the dataset size, image resolution, and hardware processing unit.

Experimental Variables

- Data Collection
  - 50 images per class
  - 250 images per class
  - 1,000 images per class

- Data Set Size for 10 classes
  - 500 samples
  - 2,500 samples
  - 10,000 samples

Experimental Variables

- Image and network resolution
  - Input shape and image resolution are the same
  - Down-sampled from an original capture resolution of 1500x1500 pixels

- Experimental results for:
  - “720p”: 736x736
  - “360p”: 384x384
  - “240p”: 256x256

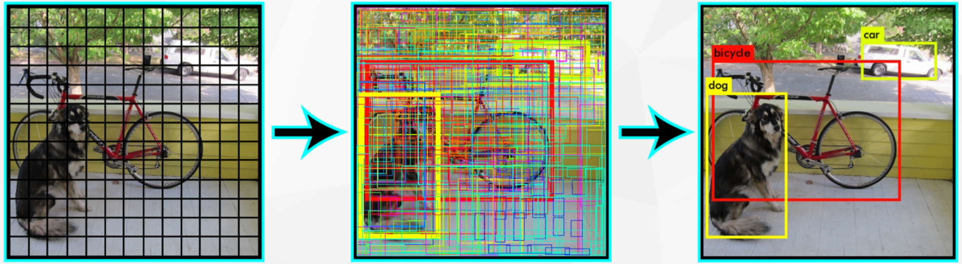
Experimental Variables

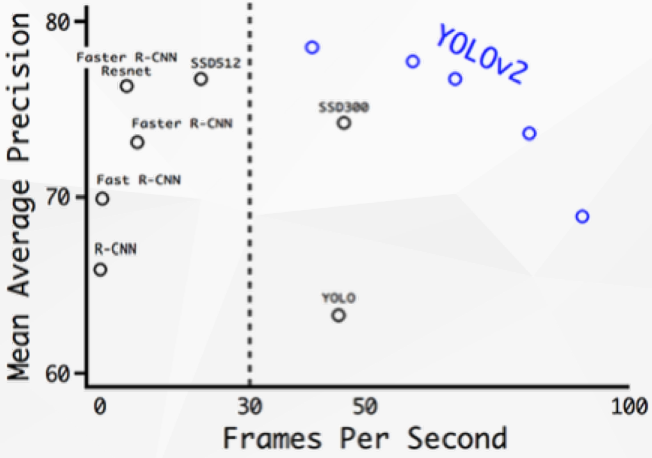
- Evaluated centralized vs. edge processing:
  - Jetson AGX Xavier Developer Kit
  - NVIDIA Tesla V100 16GB

Fixed Parameters

Fixed Parameters

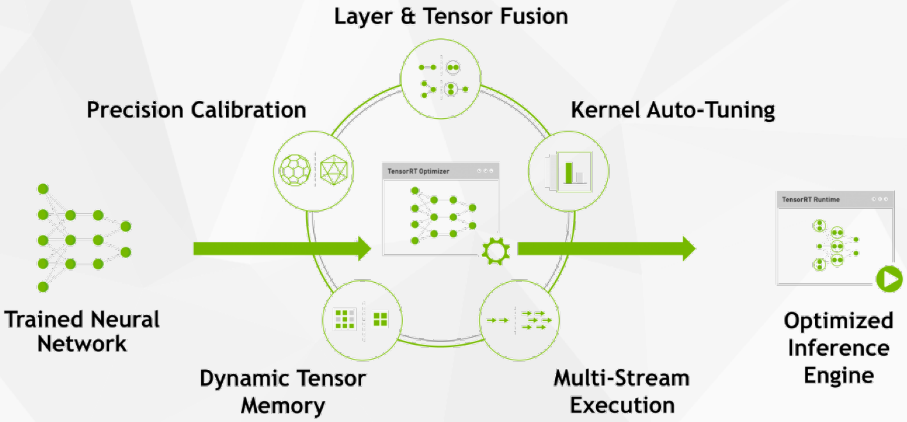
- YOLOv2
- Real-time object detection system





Fixed Parameters

- TensorRT 5.0
  - Dramatically increases inference speed
  - Small reduction in accuracy without further tweaks
  - Used INT8 precision



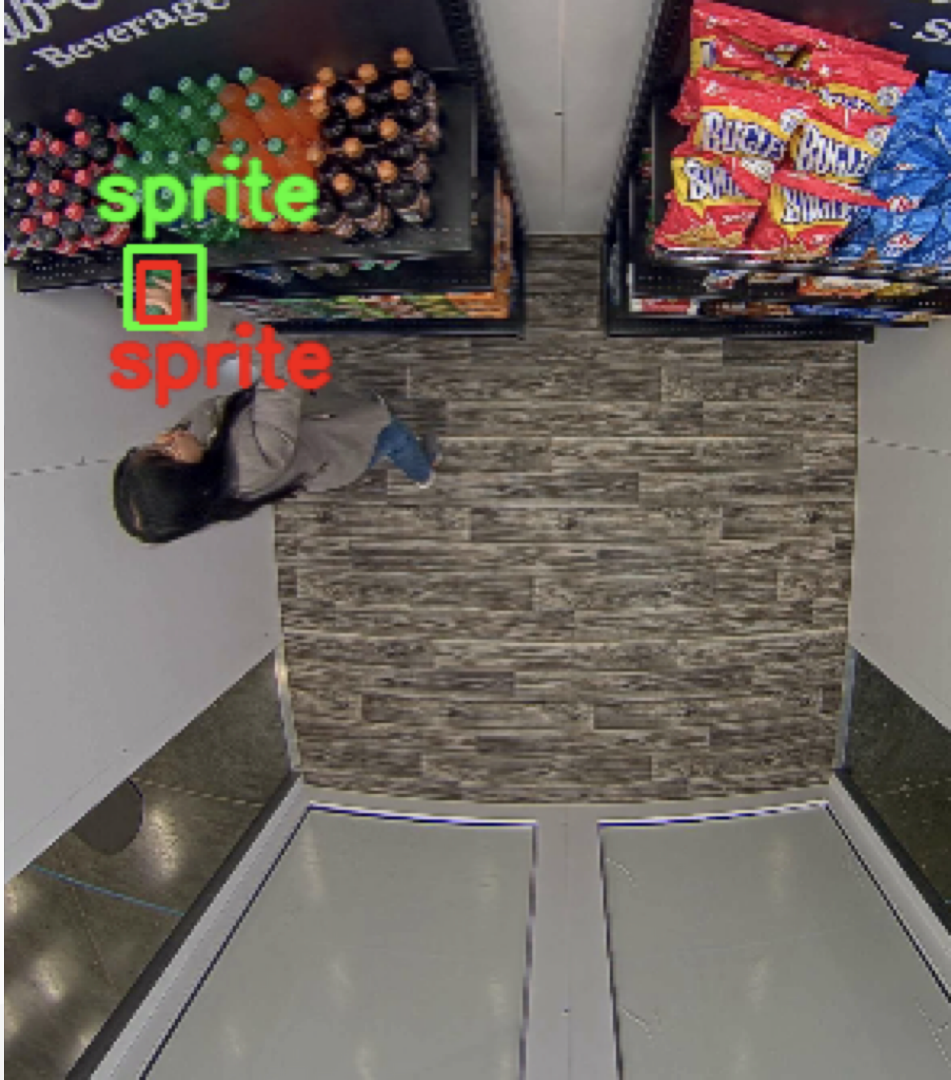
Fixed Parameters

- Mounted 9 feet high
- Axis 5MP fisheye camera
- Used 1500x1500 center patch
- 48 cameras

Visualization Reference

Ground Truth

Prediction



Validation Dataset

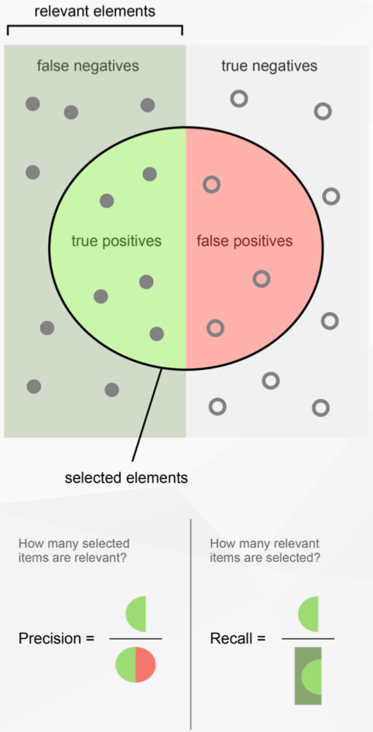
- 10 videos, 250 frames
- Small dataset for this experiment
- Due to size missing a few frames dramatically impacts metrics

Key Performance Indicators (KPI)

KPIs

- Precision:
  - How many items labeled as Sprite were Sprite.
  - Doesn't tell you about the Sprites you missed.

- Recall:
  - Out of all Sprites, how many you labeled as Sprite.
  - Doesn't tell you about 7-Up incorrectly labeled Sprite.



KPIs

- IoU of 0.5
  - Measures overlap between 2 regions.
  - How good is our prediction relative to ground truth?

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

KPIs

- How should we evaluate the trade-off between inference speed and model "accuracy"?

- It's a balance between:
  - Model can accurately detect and recognize objects but slow
  - Model does a poor job of detecting and recognizing objects but is fast

KPIs

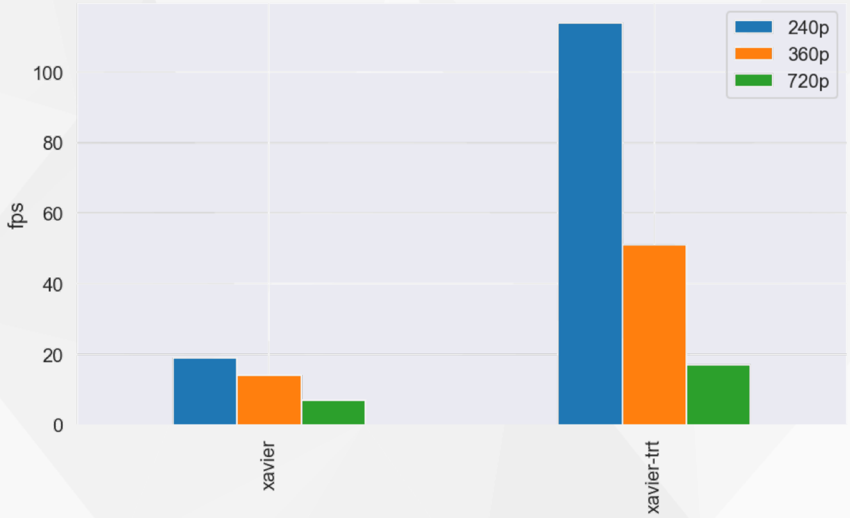
- Precision given missed frames
  - Same calculation as precision except that we penalize for missing detections in unprocessed frames
  - Account for mis-classification of items.

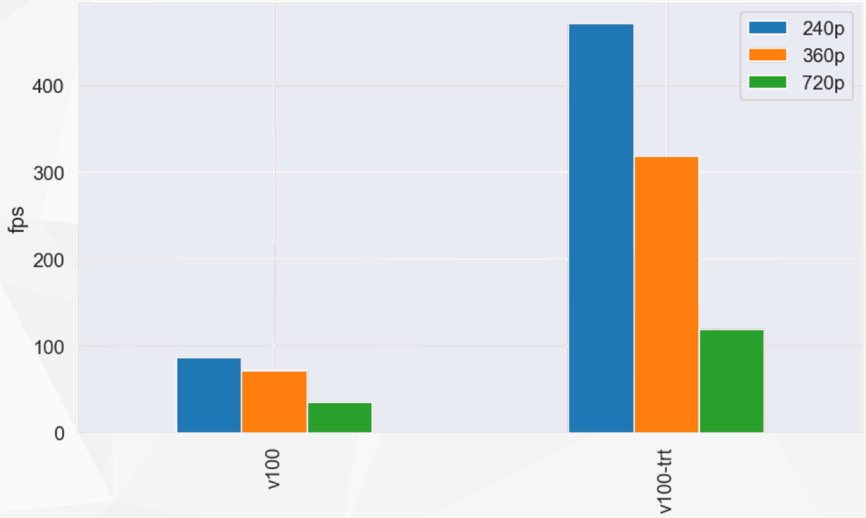
- Recall given missed frames
  - Same calculation as recall except that we penalize for missing detections in unprocessed frames
  - Account for missed detection of items.

Hardware Performance Results

TensorRT Acceleration for 1 Camera

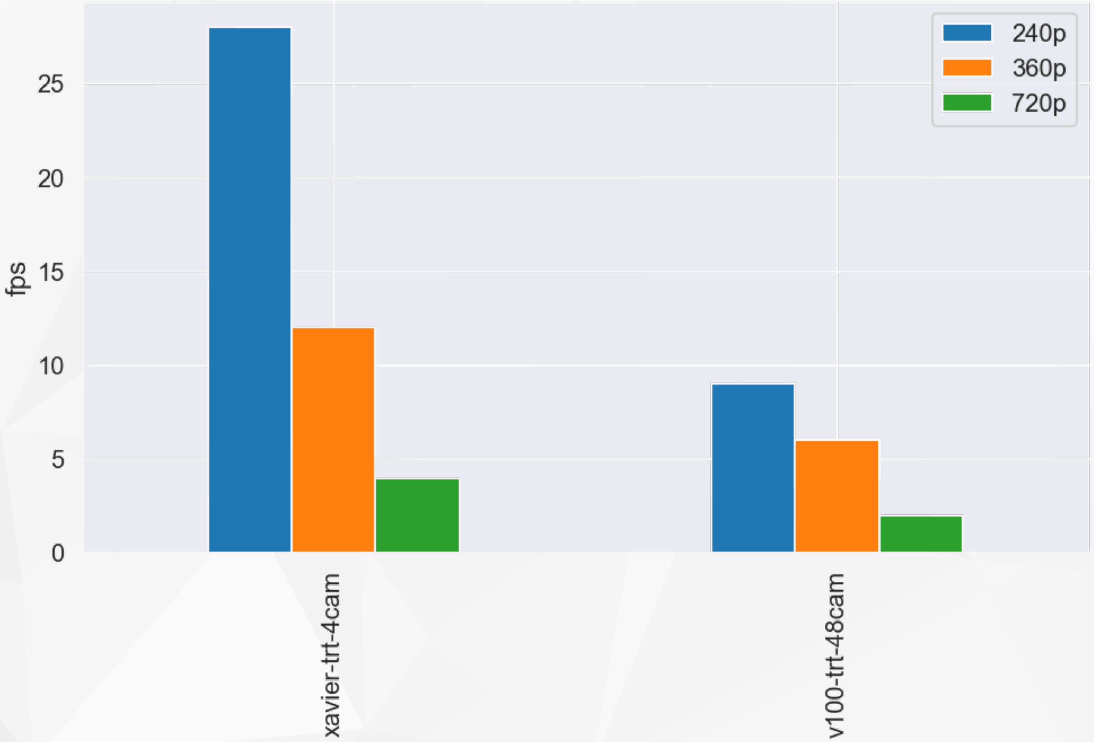
- 6x improvement in FPS
- Without TensorRT, multi-camera edge solution is infeasible.





48 Cameras: 12 Xaviers vs. V100

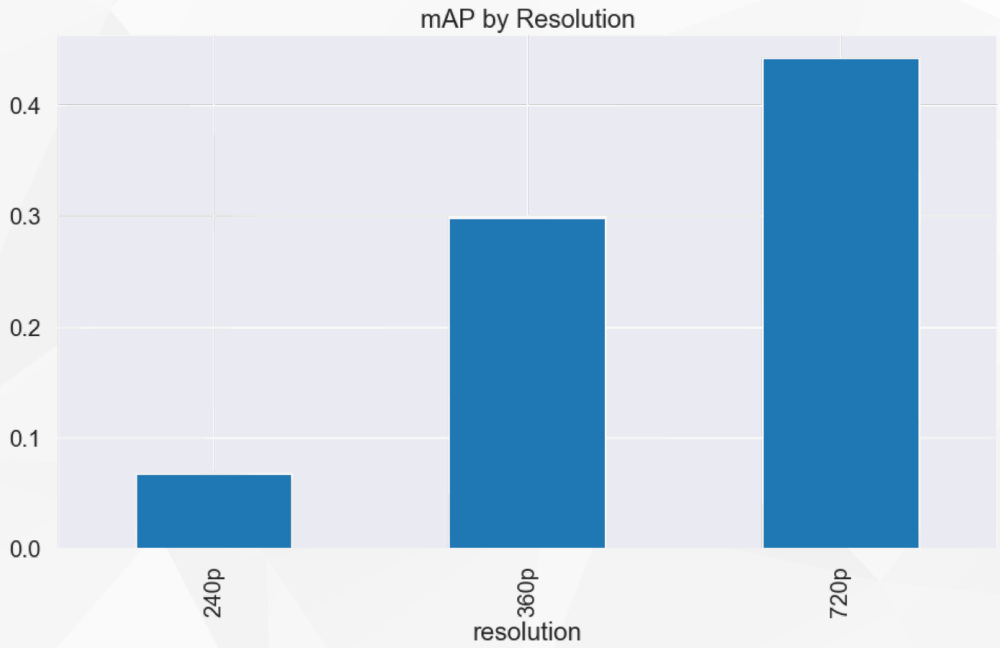
- 4 cams per Xavier
- 48 cams per V100
- Cost approx. equivalent
- < 5fps would likely be too choppy but does the data prove this?



Model Performance Results

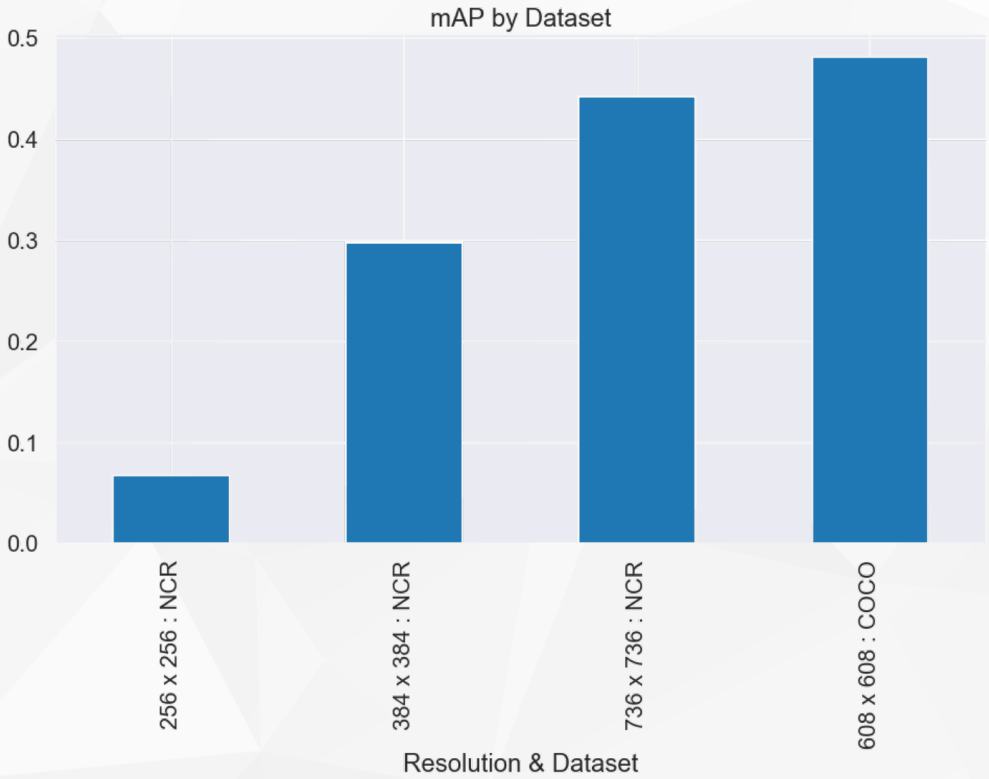
Relationship Between Resolution and mAP

- No surprise that mAP of 720p is highest
- FPS and mAP graphs look like mirror images.
- Trade-off between speed and accuracy.



YOLOv2 Performance on NCR vs. COCO Dataset

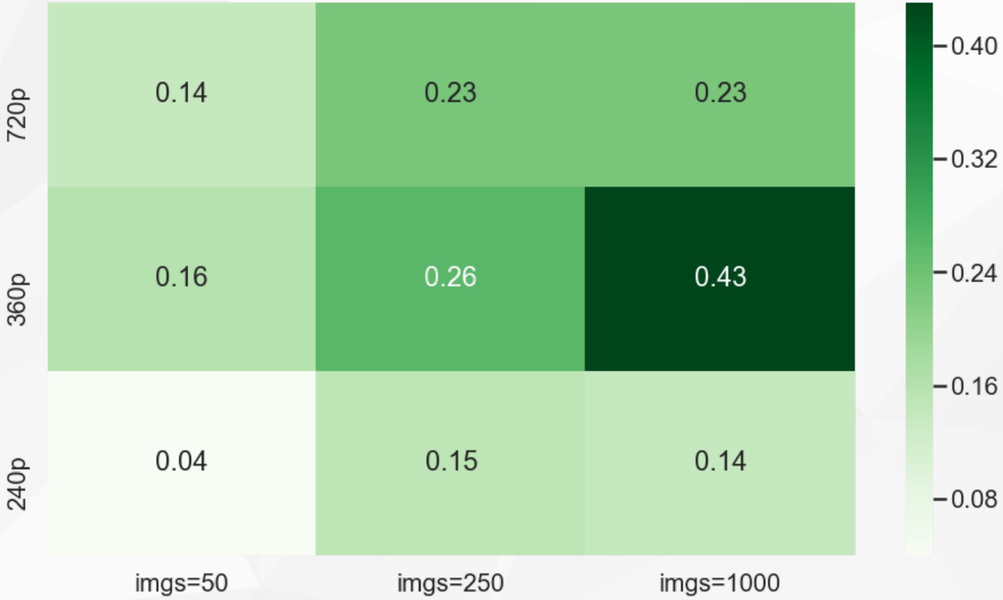
- Comparison of mAP compared to YOLOv2 trained on COCO dataset.
- COCO is a large-scale object detection, segmentation, and captioning dataset.



COCO
Common Objects in Context

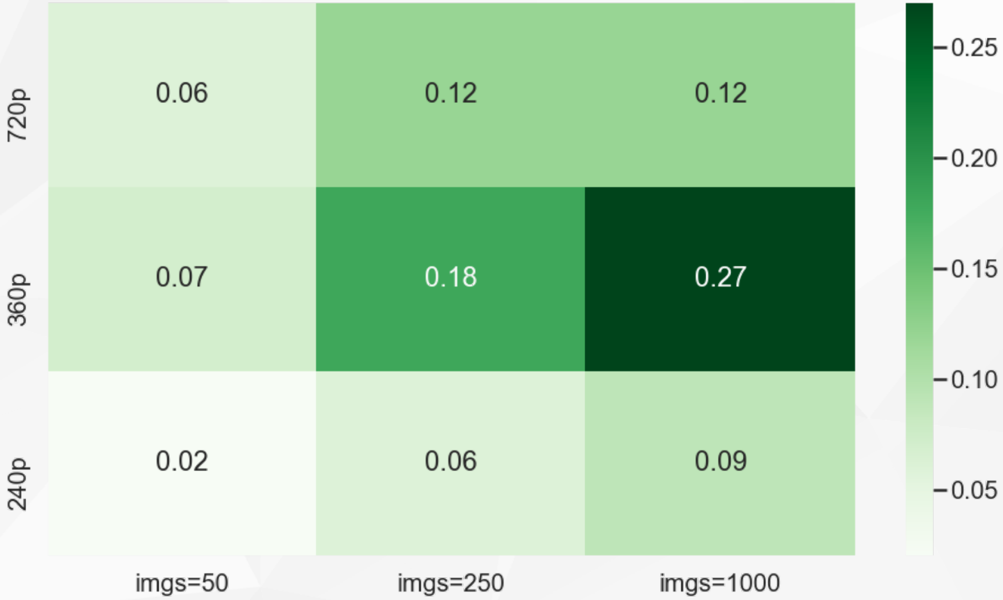
Precision Given Missed Frames on the Xavier

- Which is the best model?
- 0.43 indicates 43% as precise as the most precise model.
- Note: Values are normalized to the model with the highest mAP given no real-time constraints.



Precision Given Missed Frames on the V100

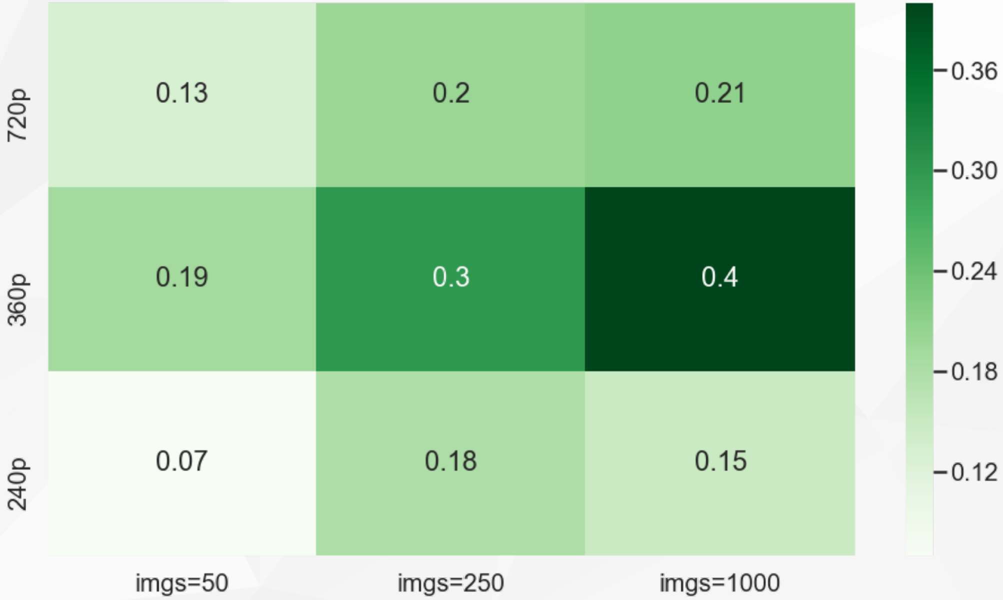
- Proportionally similar to Xavier.
- Reduced precision due to high number of cameras.



Recall Given Missed Frames on the Xavier

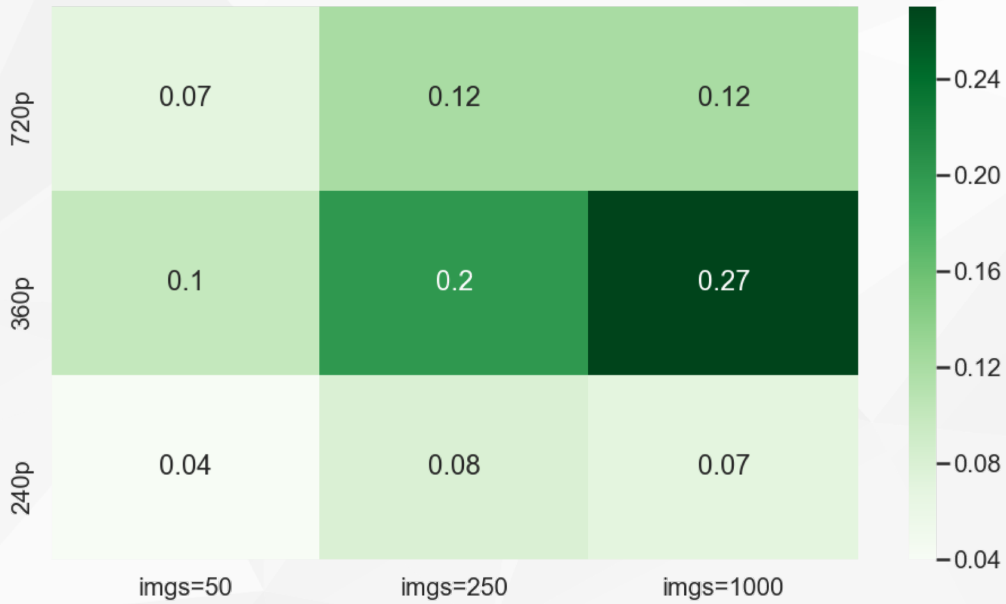
- Similar ratios to the precision results.

- Recall this is, “Out of all Sprites, how many you labeled as Sprite.”



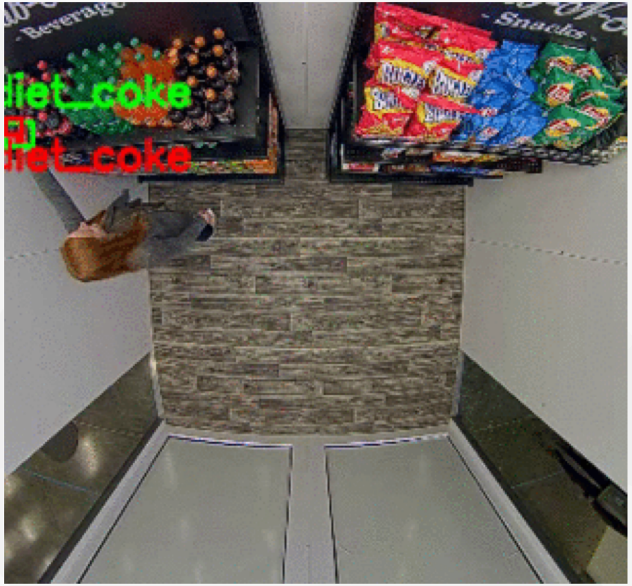
Recall Given Missed Frames on the V100

- No surprises here.

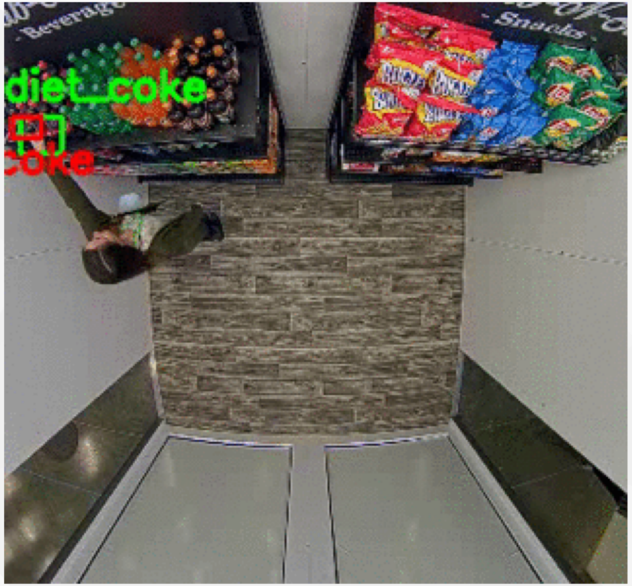


Visual Comparison

Best Xavier Model (360p / 1,000 samples)



Best V100 Model (360p / 1,000 samples)



Solution Comparison

- Decentralized wins out given similar budget.

	XAVIER	V100
DATASET SIZE	1,000 images	1,000 images
INPUT RESOLUTION	360p	360p
INFERENCE SPEED	12 fps	6 fps
REAL-TIME PRECISION	0.19	0.12
REAL-TIME RECALL	0.08	0.05

FUTURE AREAS OF RESEARCH

Opportunities for Research and Experimentation

- Larger dataset
- Multi-stage approach for localization and classification.
- Explore alternative model architectures.
- Incorporate depth.
- Sensor fusion.

- NVIDIA T4 GPU for inference.
- DeepStream SDK 3.0 or 4.0?
- Further optimize model architecture for TensorRT & GPU microarchitecture (e.g., SIDNet).

THANK YOU