KNOCK-KNOCK

A deep learning powered door camera

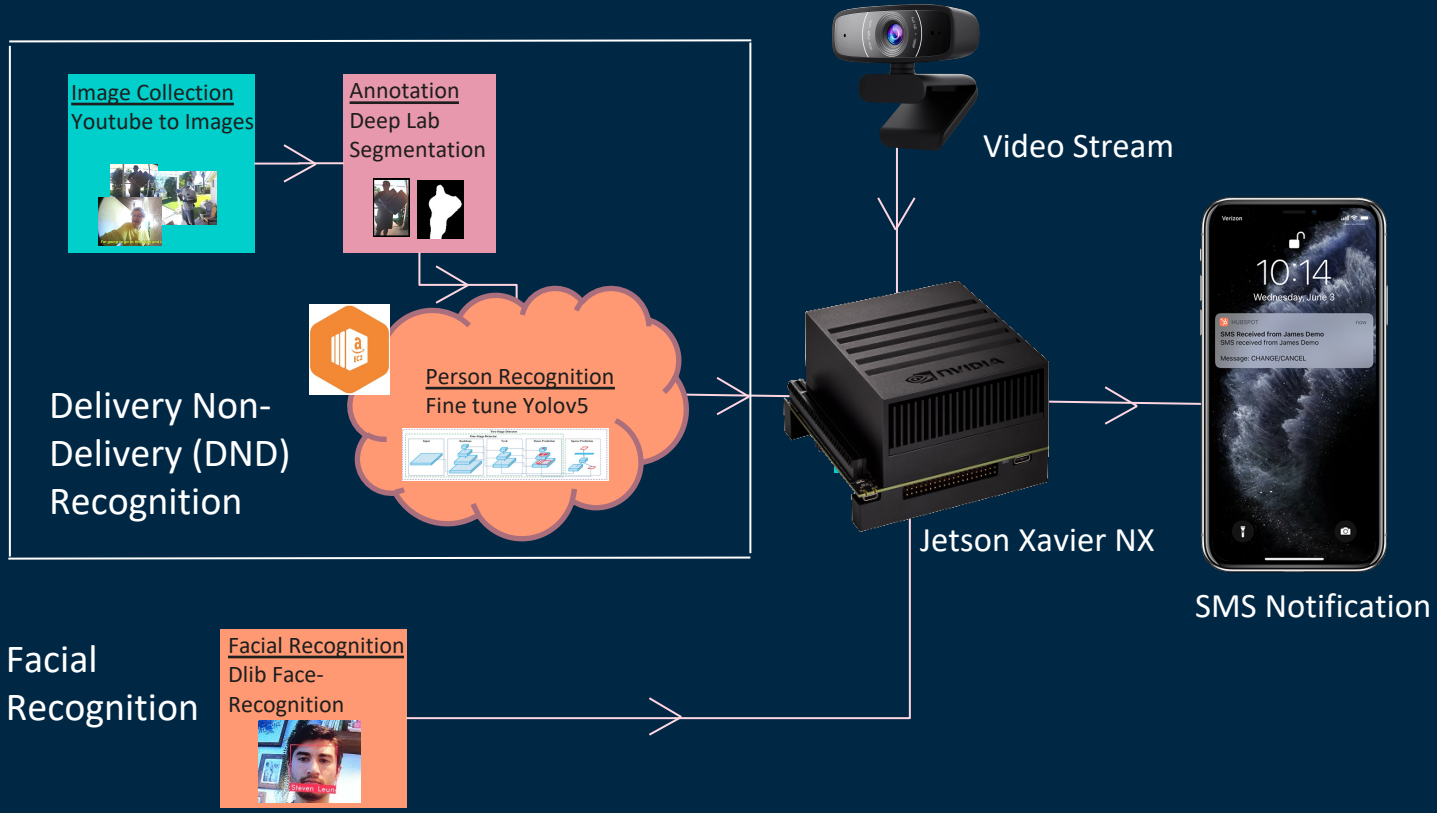
Steven Leung
Juan Ramirez
Javed Roshan

W251 Final Project Presentation

Why Knock-Knock?

- The door camera market size reached $1.83B in 2020 (16% US households)[1]
- Knock-knock is a novel door camera that recognizes common personnel at the door and notifies user:
  - Facial Recognition to identify known individuals
  - Full body recognition to identify delivery vs non-delivery personnel

Data Pipeline

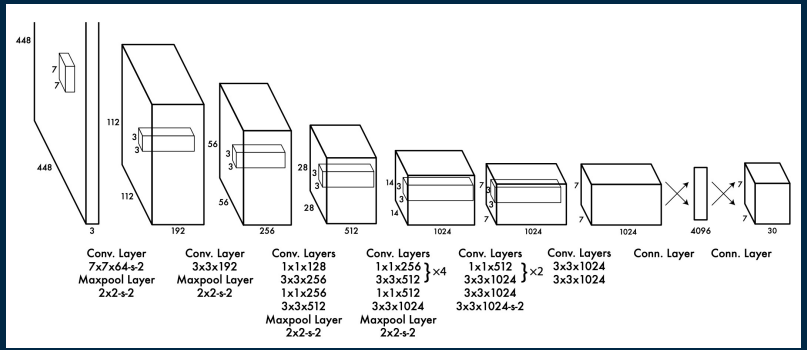


Automated Annotations- DeepLabV3 Segmentation

We used YouTube search query and DeepLab V3 (semantic segmentation) to automatically annotate 10k images from the video results.

Object Detection – YOLOv5

- Real Time Object Detection: You Only Look Once (YOLO)
- Performance compared to other models like Faster RCNN (Resnet50 based)
  52.8 FPS vs 21.7 FPS
- Architecture:
  - 24 Convolution layers with 2 FC layers
  - Pre-trained on COCO data
  - Original versions were built using Darknet framework (C based)
  - Yolov5 is written in PyTorch
- Some controversy on the use of name / transparency of model architecture details.

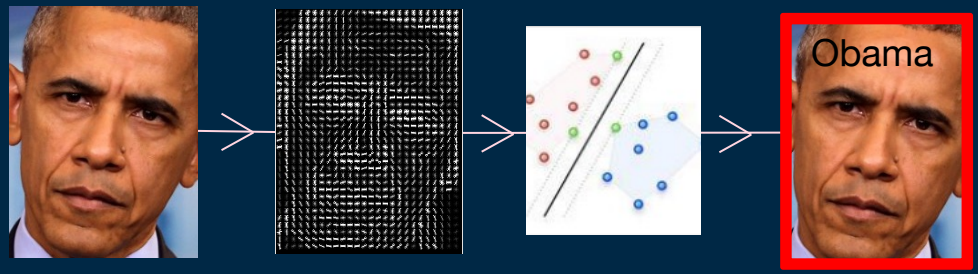


https://arxiv.org/pdf/1506.02640.pdf

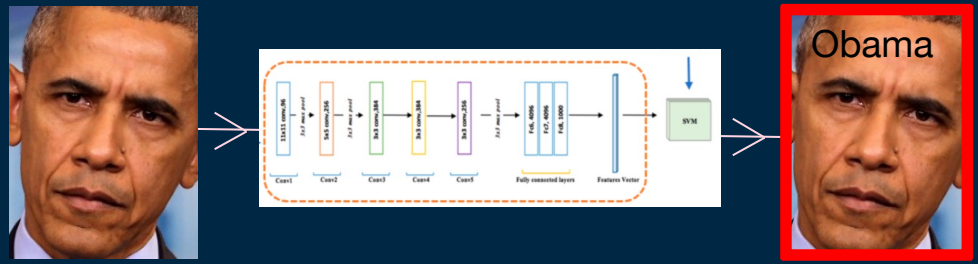
Dlib Face-Recogntion

- Dlib is a popular deep learning library written in C++ with python api.
- Face_recognition is a one-shot model utilizing Histogram Oriented Grandient (HOG) + Support Vector Machine (SVM) model or CNN based model
- Performed at 99.38% on Labeled Faces in the Wild benchmark.

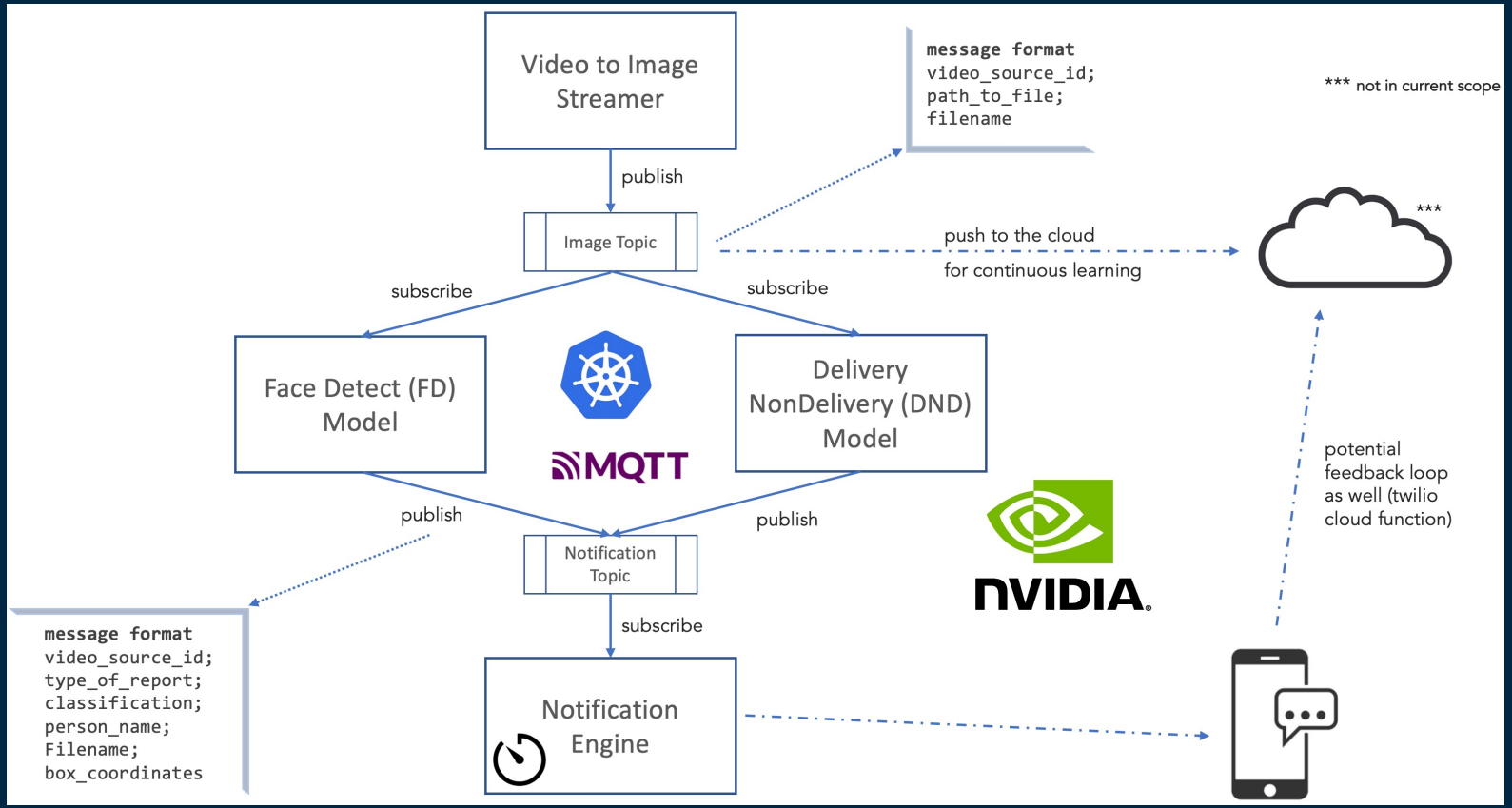
HOG + SVM



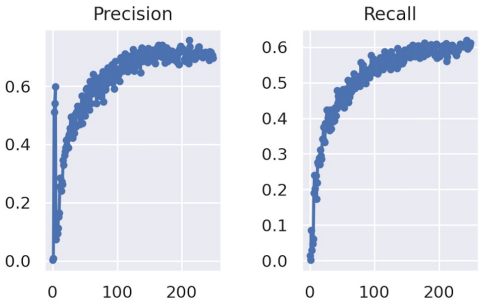
CNN

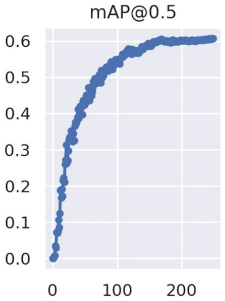


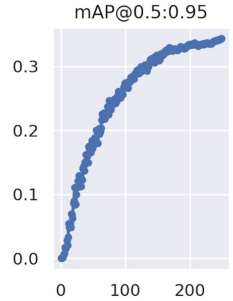
Inference Side



Model Performance

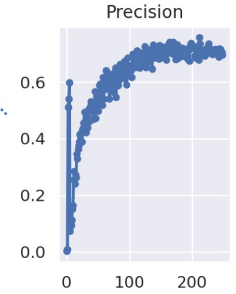


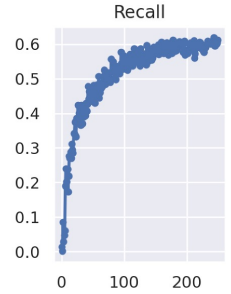


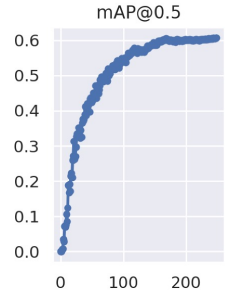


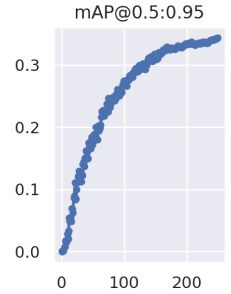
training run #2

250 epochs on AWS g4dn.xlarge
training completed in 22.423 hrs
training/valid/test: 3596/1200/1181 files
mAP@0.5 is 0.607
non-delivery: 0.598
delivery: 0.617







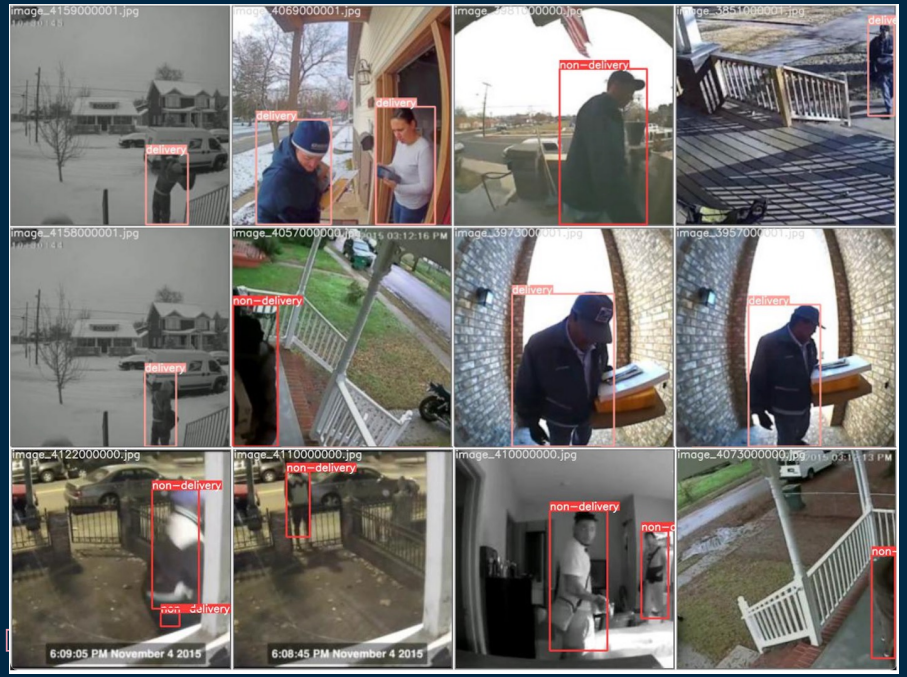


training run #1

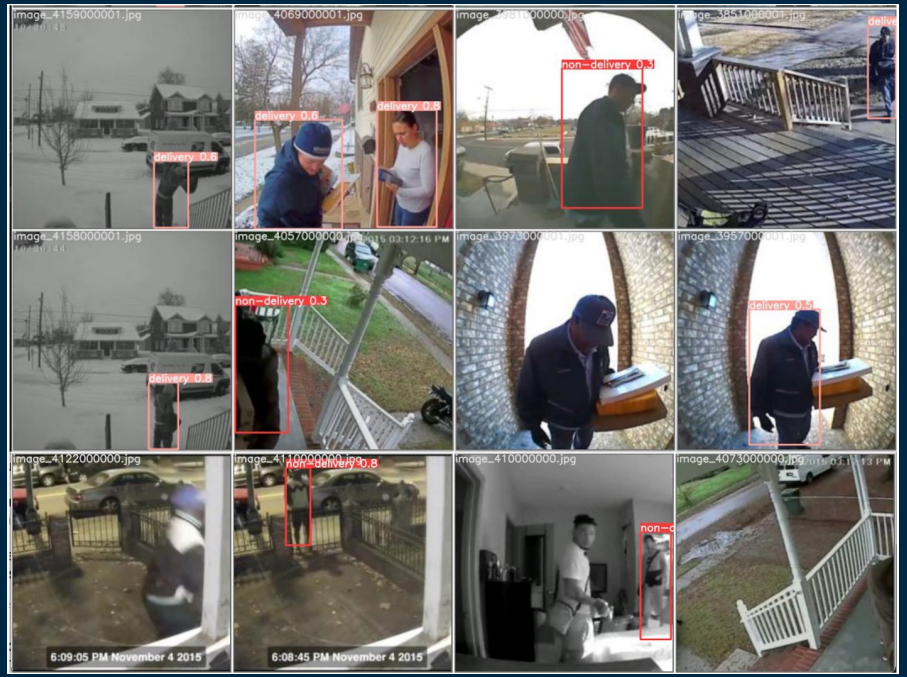
250 epochs on AWS g4dn.xlarge
training completed in 22.425 hrs
training/valid/test: 1659/575/535 files
mAP@0.5 is 0.582
non-delivery: 0.566
delivery: 0.597

Model Performance

Labelled



Predicted



Testing w/Yolov5

Testing Yolov5

- Concept of an **Event**
- 1 frame/file per second
- divided the test data into 210 events
- total image files: 1181
- model assesses each file
- max count of responses used
- classified right: 924
- classified wrong: 26
- did not classify at all: 249

78% grouped right
out of which
94% of the times the model was right
(73.32%)

Knock-Knock in Action

Challenges

YouTube queries – contained incorrect labelling and or more than one class per image adding noise into our training.

Timing & Lag – competing messages & model output synch

Notification time window logic – tuning the time window between events

Next Steps

- **User Experience**
  - Notification Enhancements
- **Continuous Training and Evaluation**
  - Allow users to provider feedback on model results for continuous improvement
- **Summarization of actions at the door**
- **Hardware**

Sources

1. https://www.grandviewresearch.com/industry-analysis/doorbell-camera-market
2. https://www.cnet.com/home/security/best-facial-recognition-security-cameras/
3. Deep Lab Segmentation https://arxiv.org/pdf/1606.00915.pdf
4. Yolov5 https://zenodo.org/record/4679653#.YQlL01NKhqsDlib
5. face rec https://face-recognition.readthedocs.io/en/latest/readme.html

Fonts & colors used

This presentation has been made using the following fonts:

Share Tech
(https://fonts.google.com/specimen/Share+Tech)

Maven Pro
(https://fonts.google.com/specimen/Maven+Pro)