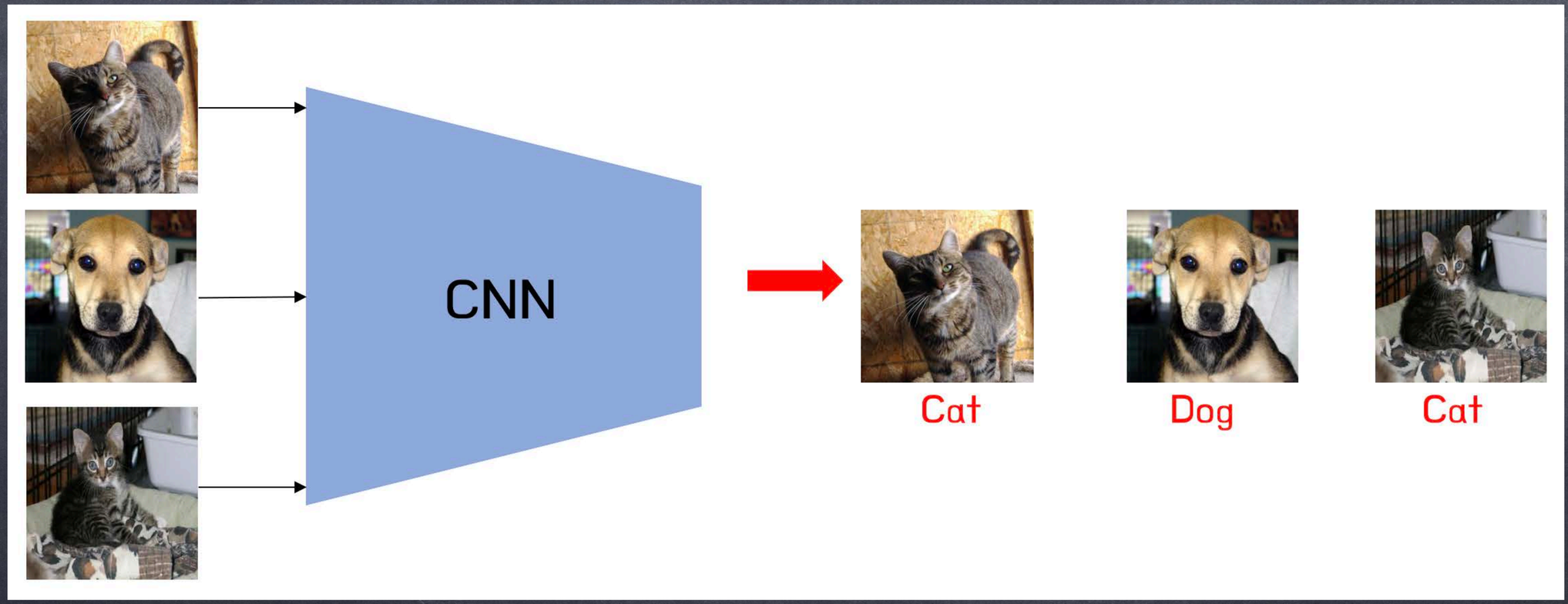
Cascade models in Computer Vision

Argo Saakyan - Computer Vision Researcher

Classification

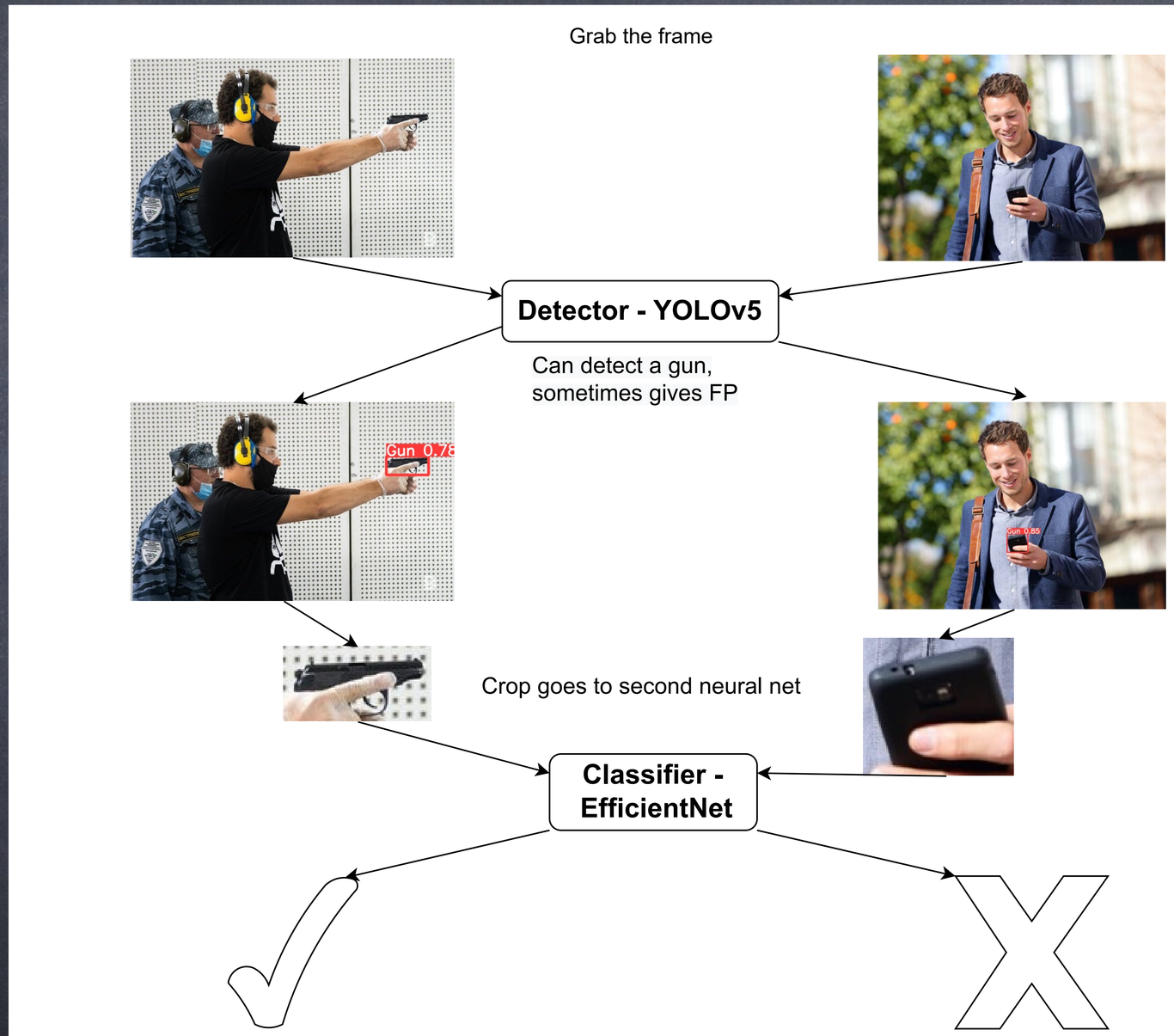


Object Detection task

Difficulties

- Need hight FPS for realtime processing
- 24/7 inference
- Accuracy/speed trade-off

Cascade example



Why does it work?

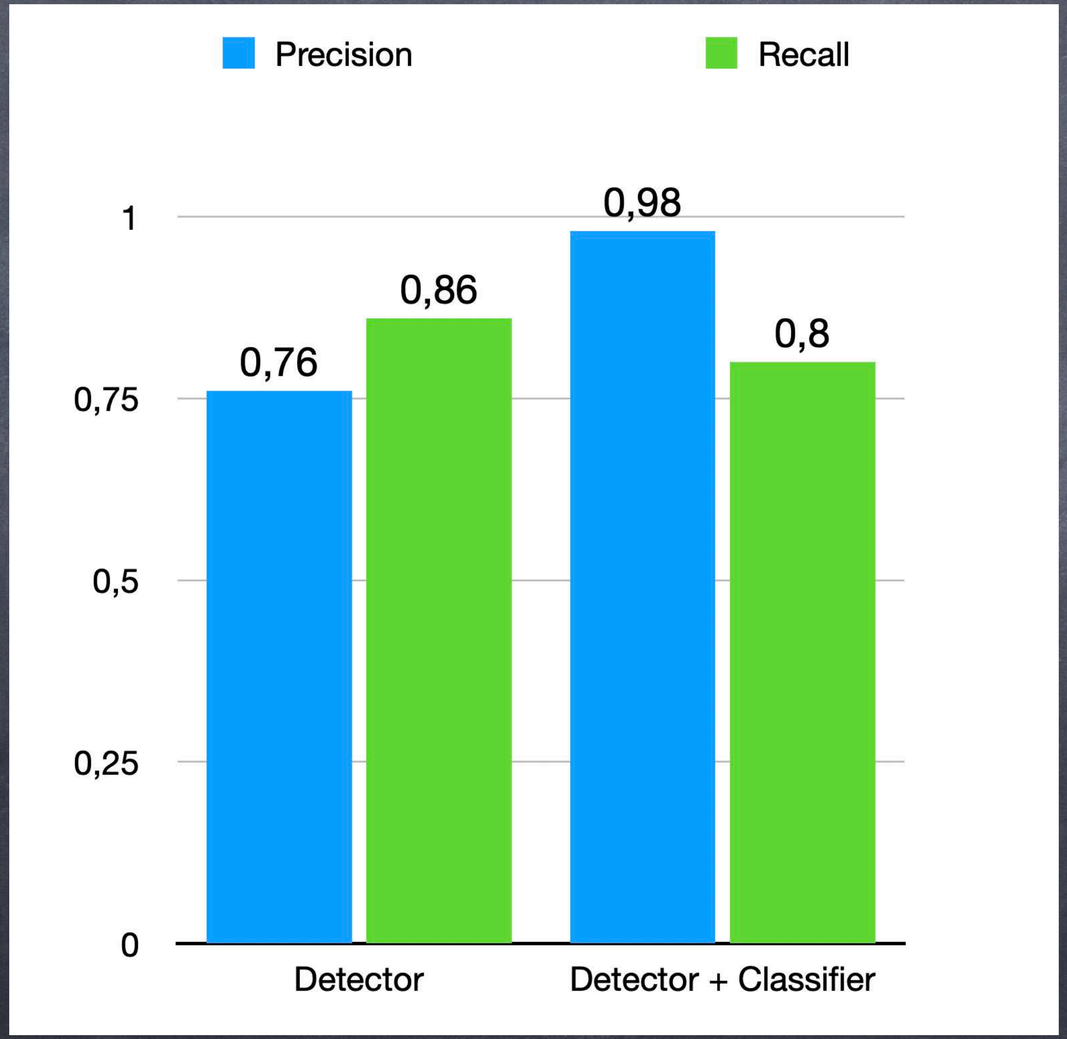
Speed:

- Speed of the solution = speed of the detector
- Can choose fast detector and not lose accuracy

Accuracy:

- Ensemble
- Higher res validation
- Different data
- Best of both worlds
- Retraining

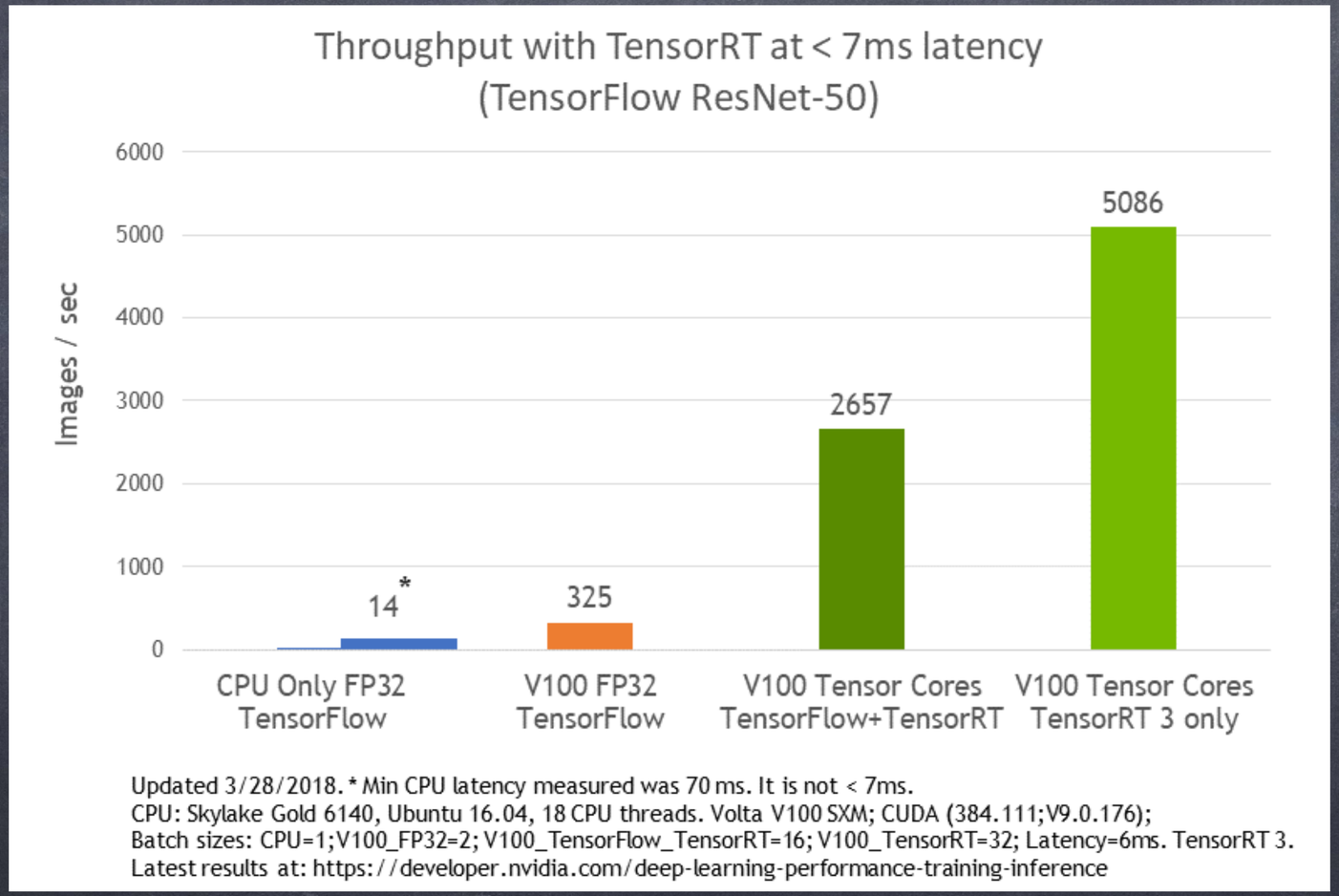
Metrics


Precision
Recall
1
0,75
0,5
0,25
0
0,76
0,86
0,98
0,8
Detector
Detector + Classifier

Dataset

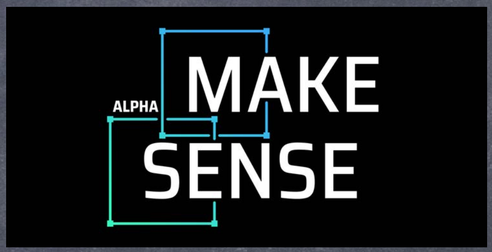
- Detector - close to inference
- Classifier - crops from detector

Deployment


Throughput with TensorRT at < 7ms latency
(TensorFlow ResNet-50)
Images / sec
6000
5000
4000
3000
2000
1000
0
14*
325
2657
5086
CPU Only FP32 TensorFlow
V100 FP32 TensorFlow
V100 Tensor Cores TensorFlow+TensorRT
V100 Tensor Cores TensorRT 3 only
Updated 3/28/2018. * Min CPU latency measured was 70 ms. It is not < 7ms.
CPU: Skylake Gold 6140, Ubuntu 16.04, 18 CPU threads. Volta V100 SXM; CUDA (384.111;V9.0.176);
Batch sizes: CPU=1;V100_FP32=2; V100_TensorFlow_TensorRT=16; V100_TensorRT=32; Latency=6ms. TensorRT 3.
Latest results at: https://developer.nvidia.com/deep-learning-performance-training-inference

Examples

- YOLOv8 - Detector
- Makesense.ai - Labeling
- EfficientNet - Classifier
- Deploy with Triton Server



Thank you!

- linkedin.com/in/argo-saakyan/