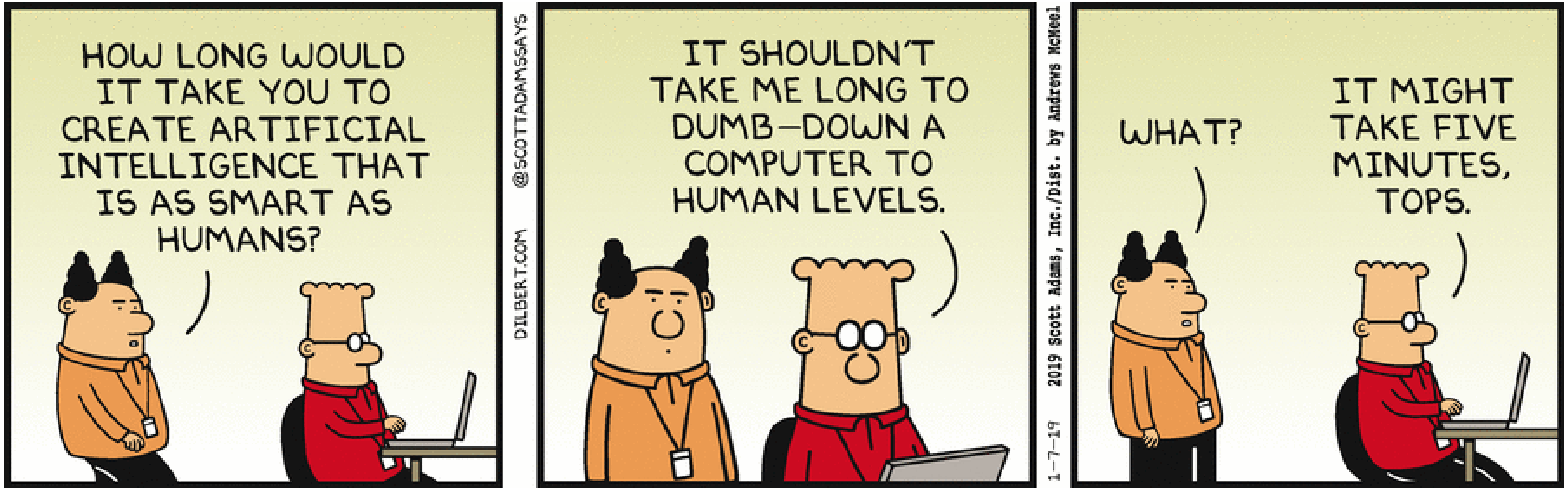
embedded VISION summit

Accelerating the Creation of Custom, Production-Ready AI Models for Edge AI
NVIDIA Tools, Part 1

Akhil Docca
Product Marketing, NVIDIA Corporation

Building an AI Application Is Hard


HOW LONG WOULD IT TAKE YOU TO CREATE ARTIFICIAL INTELLIGENCE THAT IS AS SMART AS HUMANS?
IT SHOULDN'T TAKE ME LONG TO DUMB-DOWN A COMPUTER TO HUMAN LEVELS.
WHAT?
IT MIGHT TAKE FIVE MINUTES, TOPS.
DILBERT.COM @SCOTTADAMSSAYS
1-7-19 © 2019 Scott Adams, Inc./Dist. by Andrews McMeel

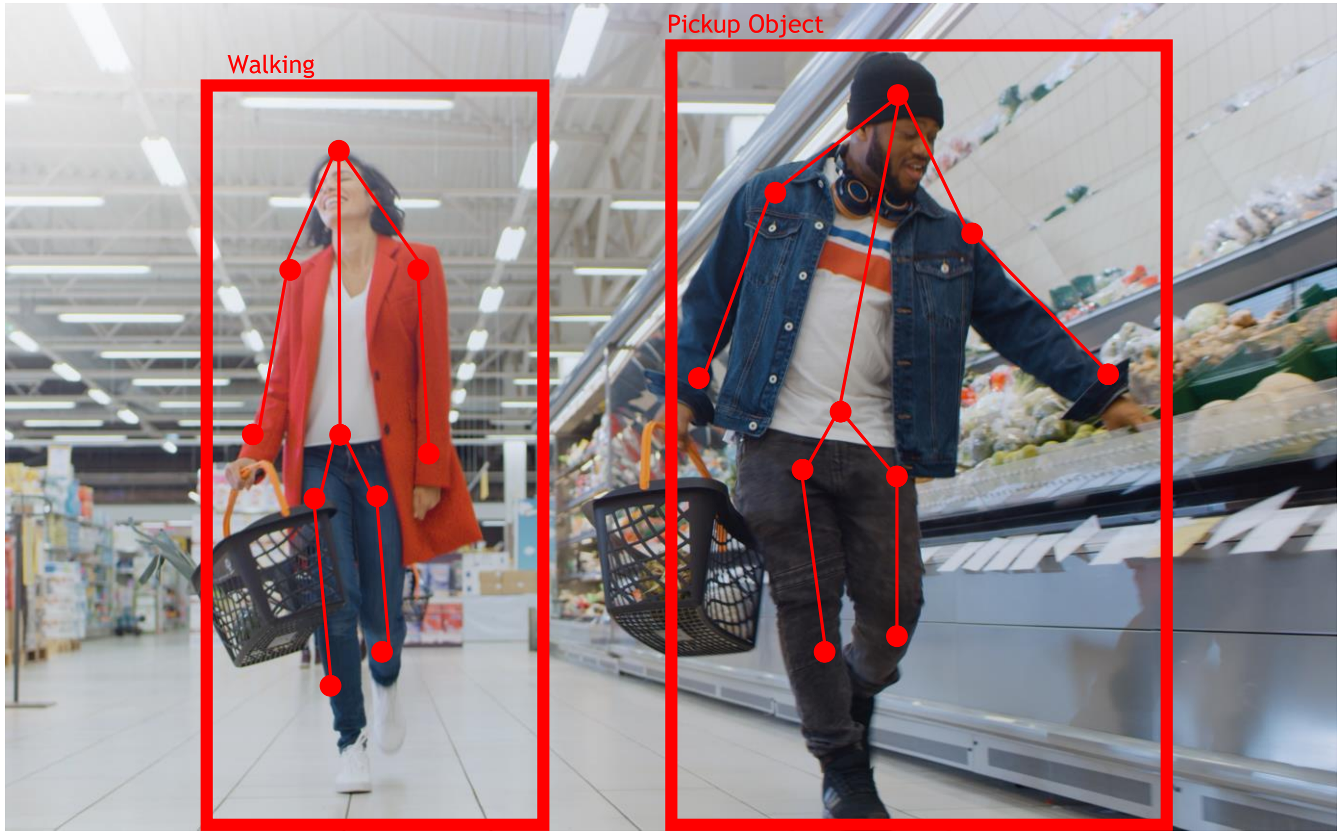
Core Elements of an Autonomous Shopping App

Customer enters the store

Picks up the item(s) of choice

Leaves the store and gets billed automatically

Core Elements of an Autonomous Shopping App



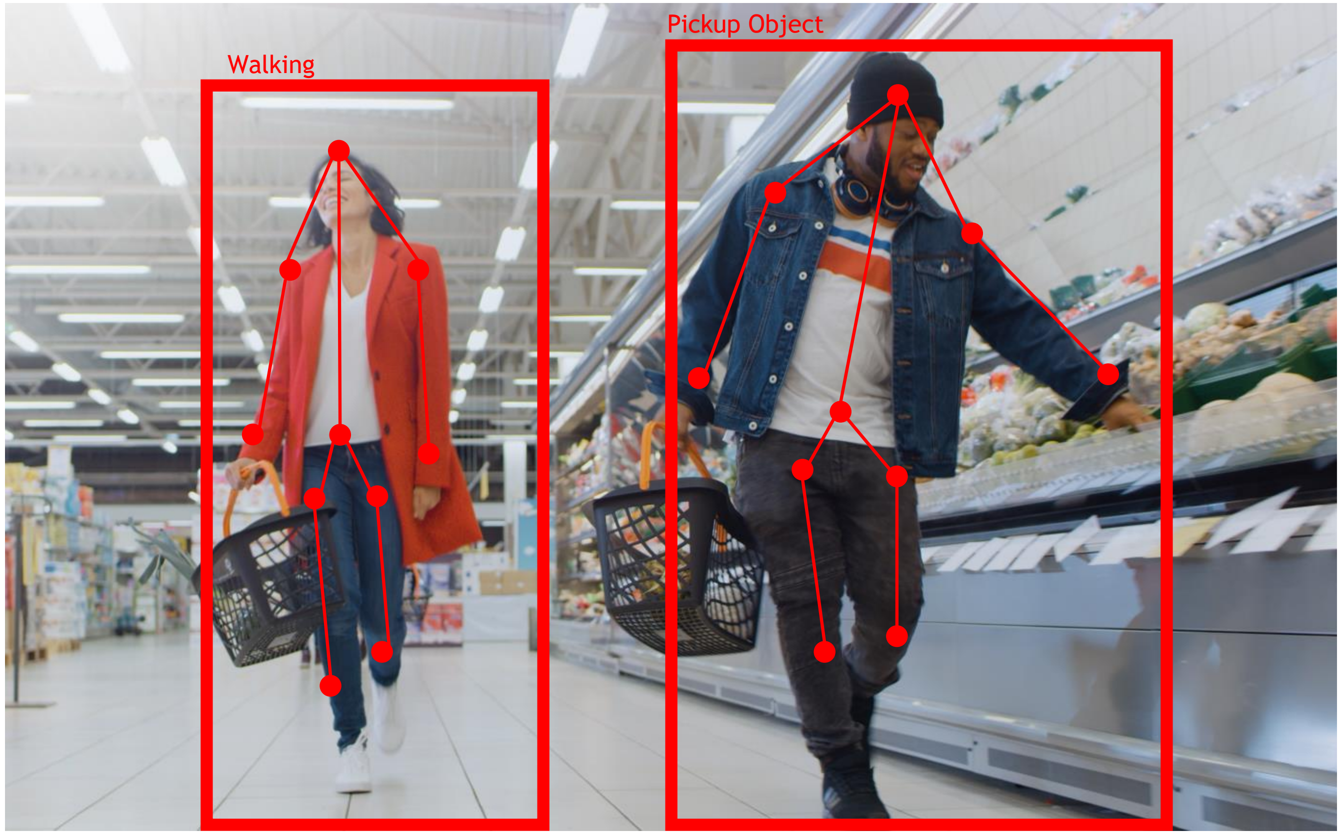
Detect person

Detect pose based on key points

Recognize actions

Identify the item(s) picked

Core Elements of an Autonomous Shopping App



OPTION 1

Build and train a model from scratch

OPTION 2

Customize a model from a model zoo

Option 1: Training from Scratch

First 5000 Days NFT (Image Courtesy: The New York Times)

PeopleNet – An NVIDIA built people detection model

Data

- 3.5M images
- 16M+ people in the images
- 40 people, 5 years to collect, curate and label

Training and Optimizing

- 30 Days A100 8 GPUs
- Several months before reaching production-quality

Option 2: Train from Existing Models

Accuracy? Dataset? Performance?

Human Detection

Regional-based Convolutional Neural Networks (R-CNN)

Visual People Detection

People Detector

Custom People Detection

Person Detection Retail

People Detection

Pedestrian Counting

Multiple Human Detection

You Only Look Once (YOLO)

YOLO

Faster R-CNN

Object Detection

Pedestrian Detection

Single Shot Detector

Write Code, Train, Iterate, Customize and Optimize

Caffe2

Chainer

K

mxnet

PaddlePaddle

PyTorch

TensorFlow

Still requires lots of expertise and time!

Industry Wide Challenge

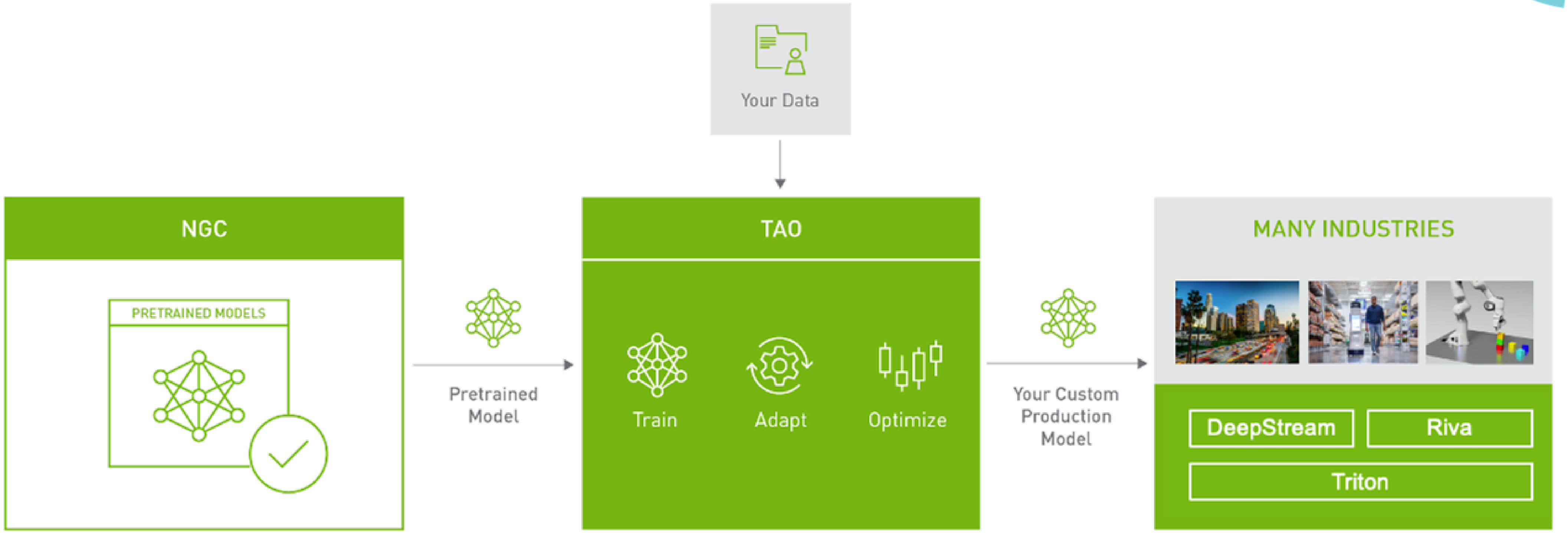
LACK OF DATA

SKILLED PEOPLE

IS THERE AN OPTION 3?

1) O'REILLY SURVEY OF MORE THAN 3000 DEVELOPERS (2021)
2) NVIDIA SURVEY OF MORE THAN 1000 ENTERPRISES (2021)

The NVIDIA TAO Framework



TRAIN EASILY

Easy to use solutions that abstract away the AI framework complexity

CUSTOMIZE FASTER

Fine tune NVIDIA pre-trained AI models with fraction of the data as opposed to training from scratch

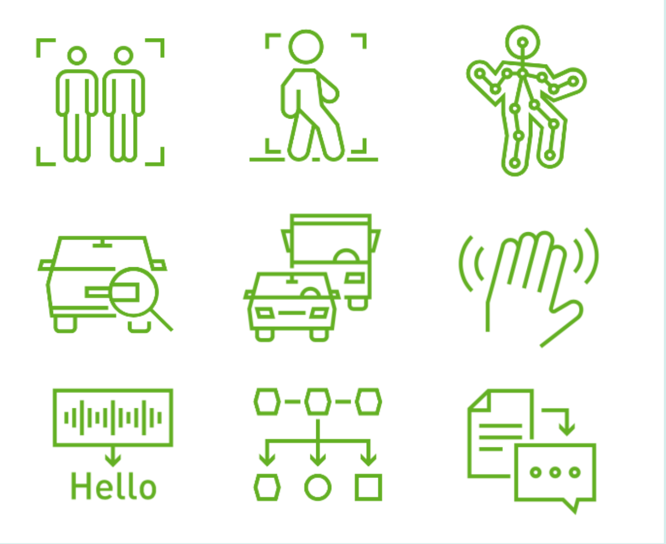
OPTIMIZE FOR DEPLOYMENT

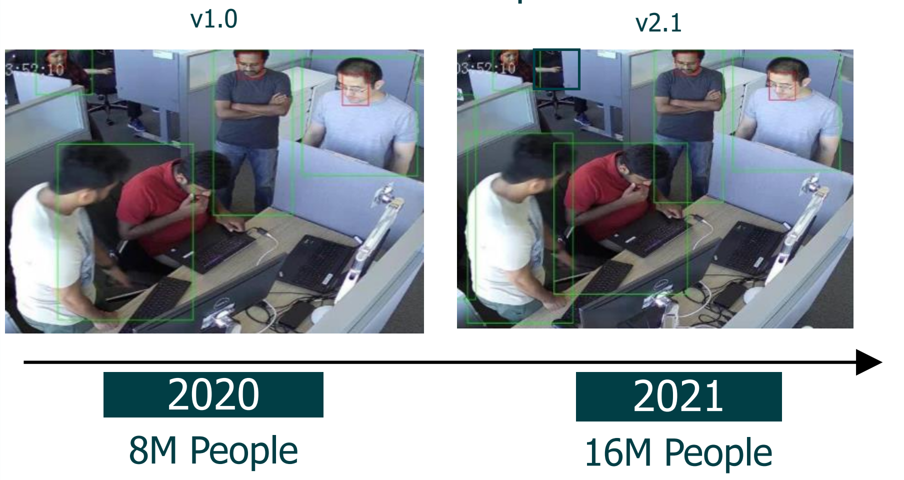
Optimize for low latency and high-throughput

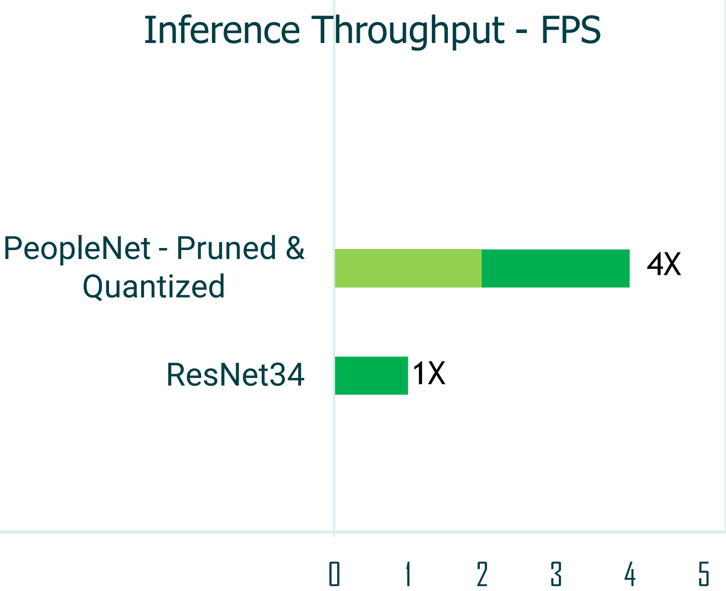
INTEGRATE AND DEPLOY

Integrate the optimized models from TAO into DeepStream (Vision) and Riva (Speech)

The Power of Pretrained Models




Evolution of the PeopleNet Model


Inference Throughput - FPS

WIDE RANGE OF USE CASES

100+ permutations of NVIDIA-optimized model architectures (EfficientDet, YOLOv3/v4)

Task based models - People Detection, Vehicle Detection, Gaze, Speech Recognition and Text to Speech

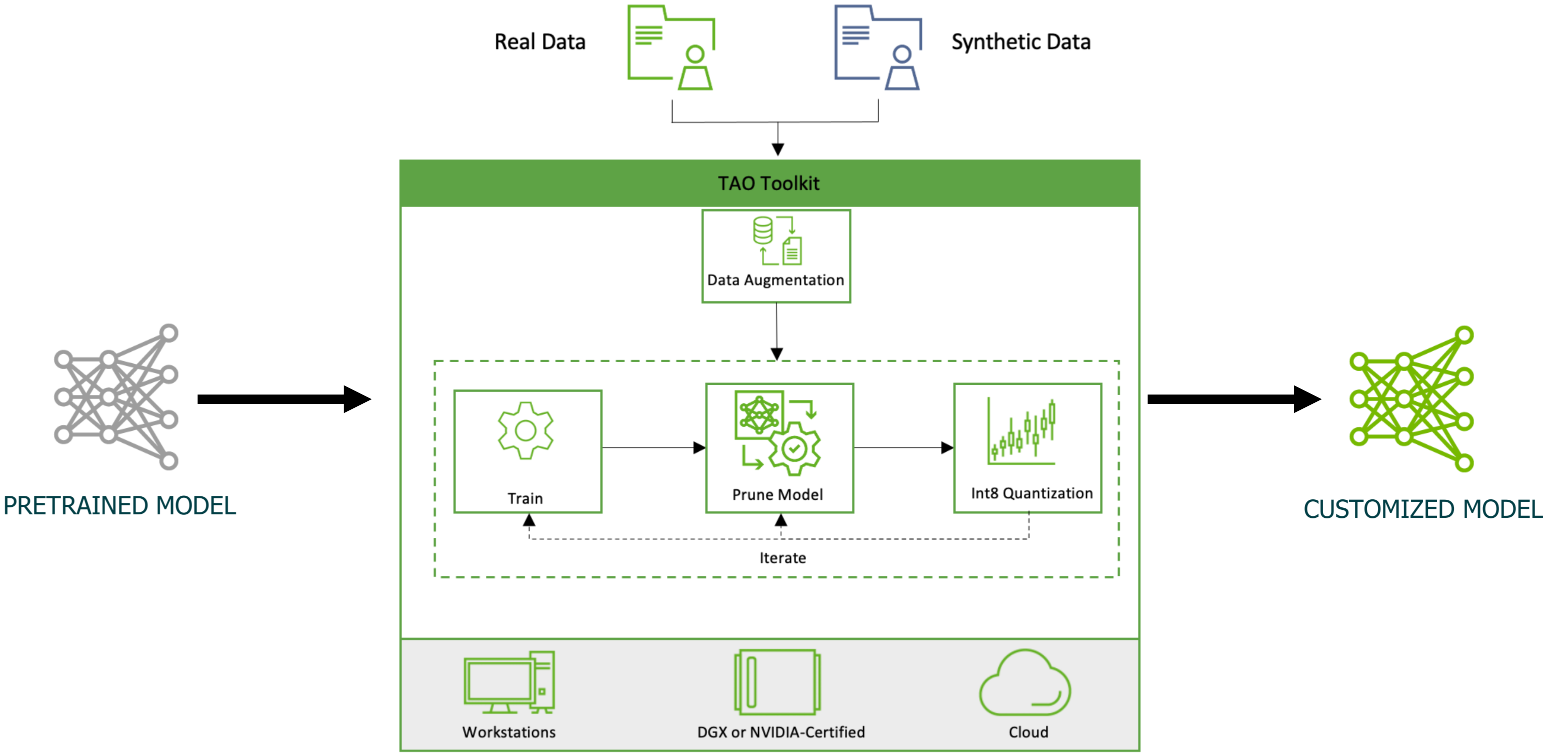
HIGHLY ACCURATE

Trained and continuously updated by experts so you can adapt to your domain or deploy as-is

OPTIMIZED FOR INFERENCE

Deploy in the data center or at the edge

A Closer Look at the NVIDIA TAO Toolkit



RUN ANYWHERE

Container based solution

TURNKEY JUPYTER NOTEBOOKS

Computer Vision, Speech and NLU

DATA AUGMENTATION

Spatial and color transformation

MULTI-GPU AND MULTI-NODE

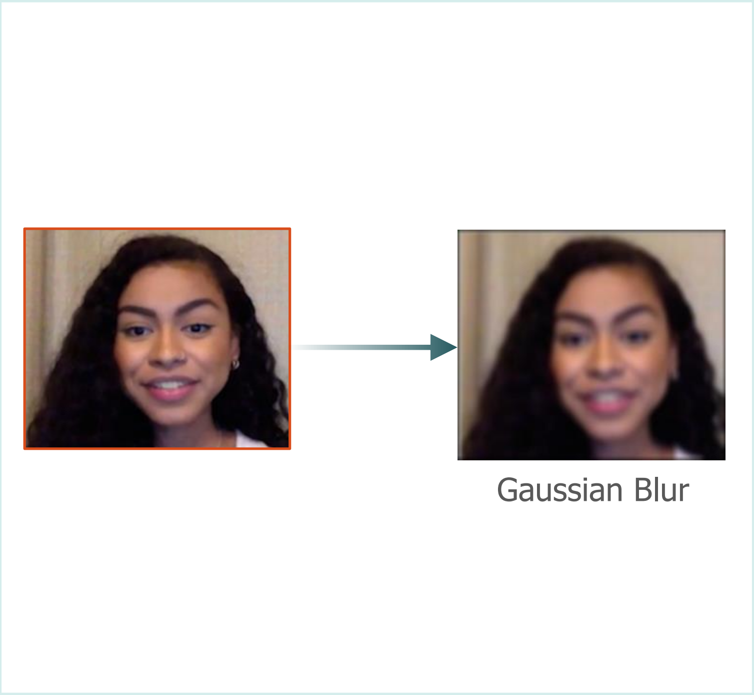
Accelerated model training

PRUNE AND QUANTIZE AWARE TRAINING

More than 4X speed up in inference

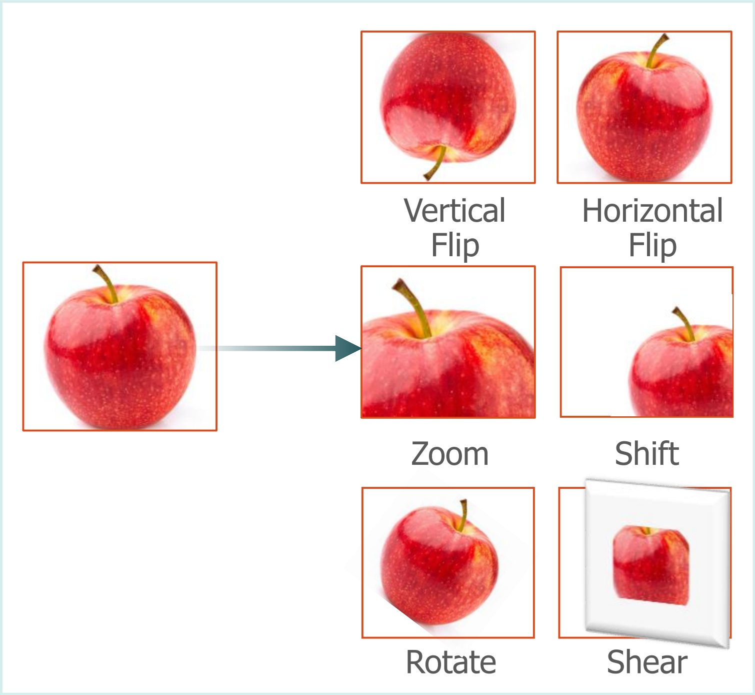
Data Augmentation

BLUR



Gaussian Blur

SPATIAL



Vertical Flip

Horizontal Flip

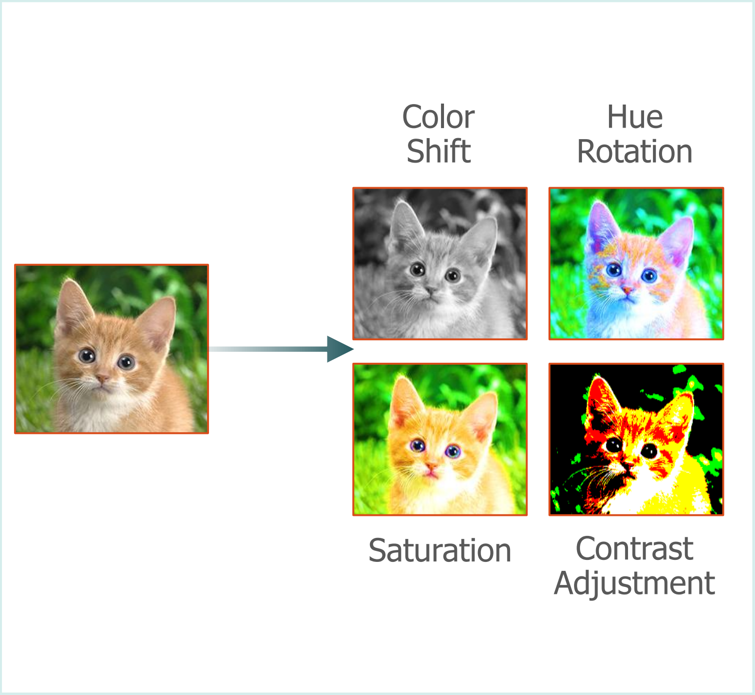
Zoom

Shift

Rotate

Shear

COLOR



Color Shift

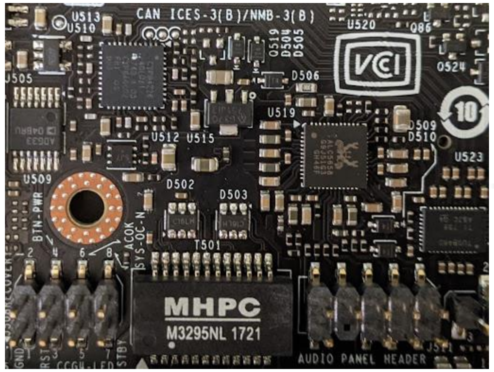
Hue Rotation

Saturation

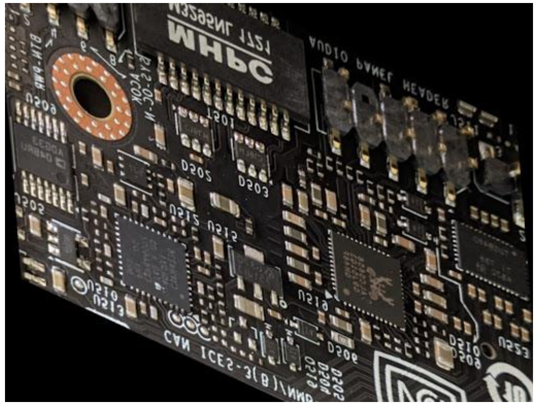
Contrast Adjustment

Data Augmentation

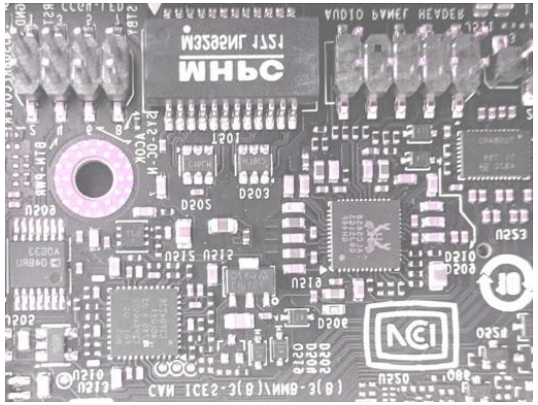
ORIGINAL

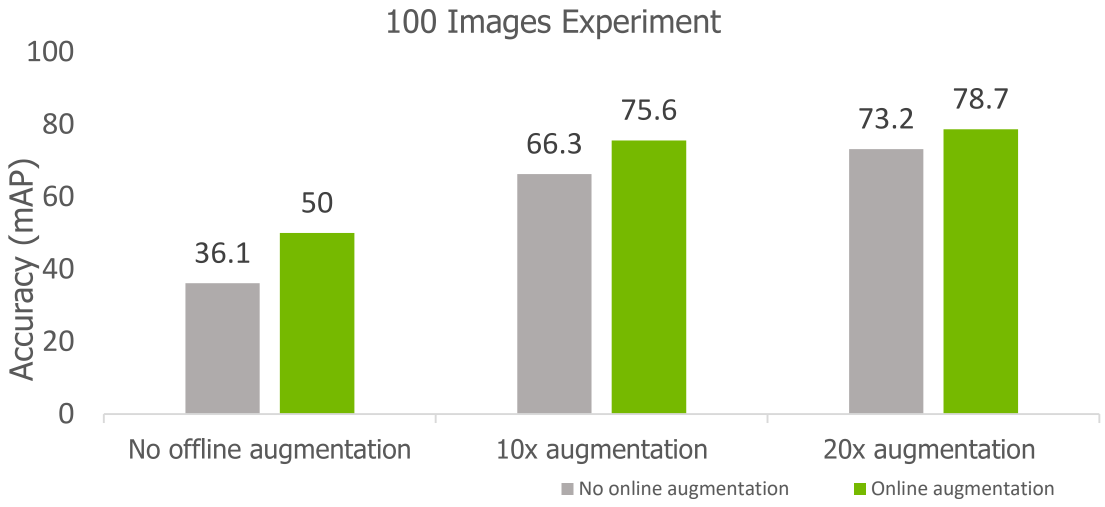


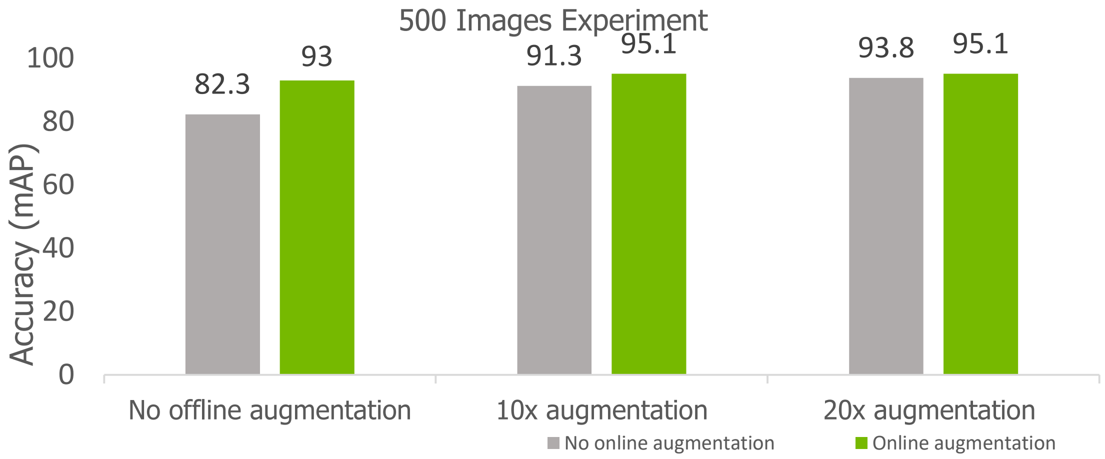
ROTATED



BRIGHTNESS ADJUSTED







TAO Toolkit White Paper

Git Hub Project Repository

Art of the Possible – Camera Types

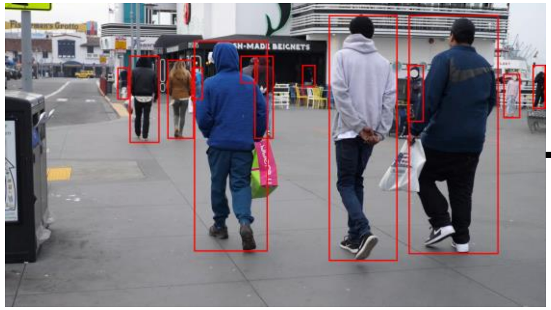
Thermal IR Data Set

TAO TOOLKIT

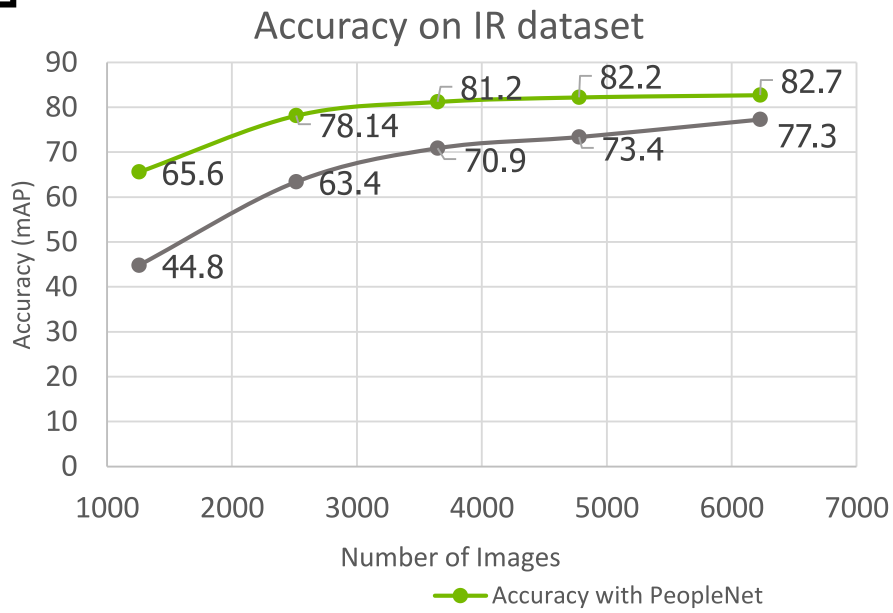
Train

Adapt

Optimize



PeopleNet Model

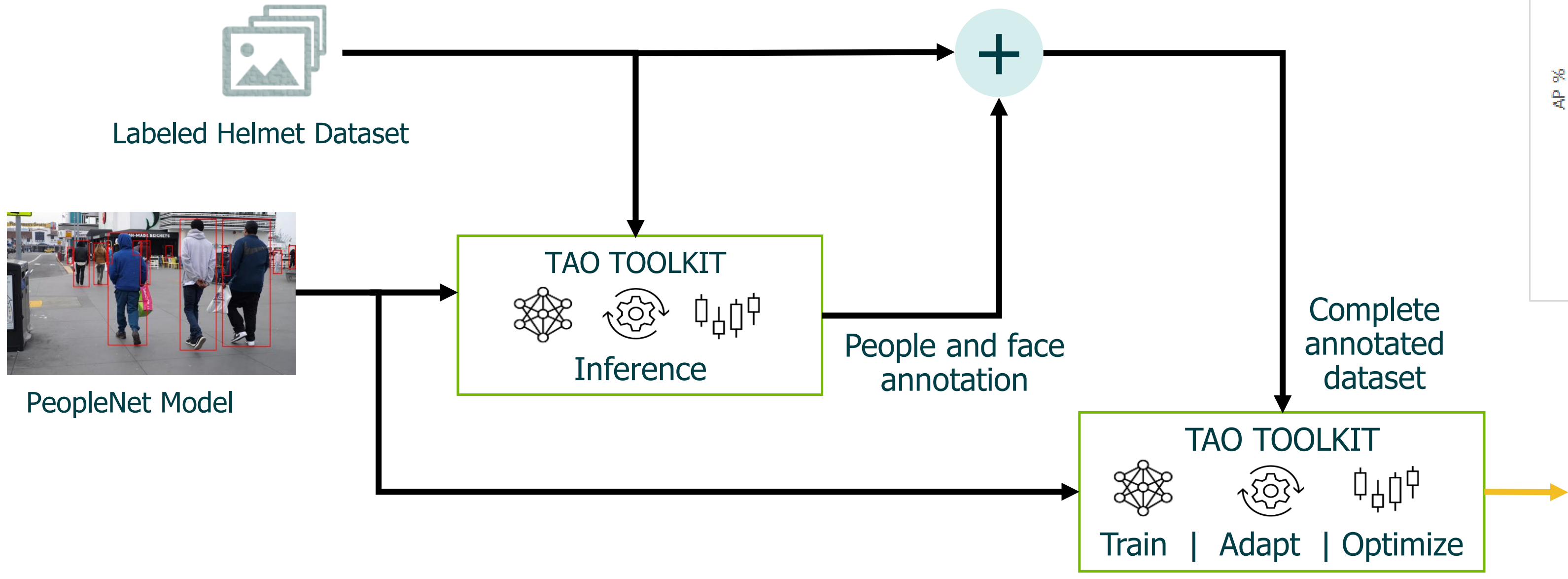


TAO Toolkit White Paper

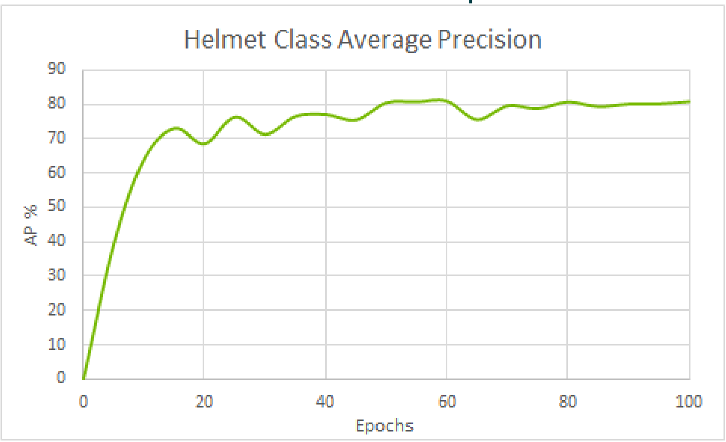
Git Hub Project Repository

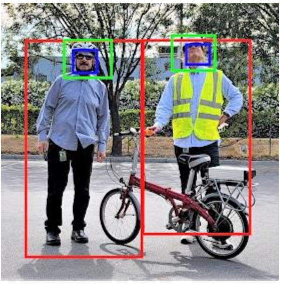
Art of the Possible – Adding New Classes

Goal: Add "Helmet" class to existing People detect model



80% AP over 100 epochs





Model with Helmet, People and Face Detection

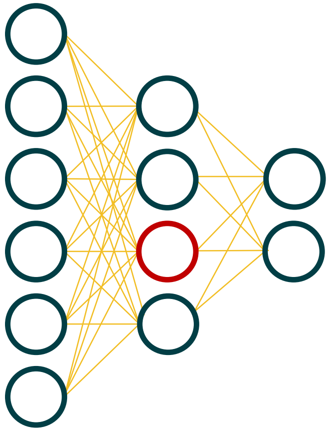
TAO Toolkit White Paper

Git Hub Project Repository

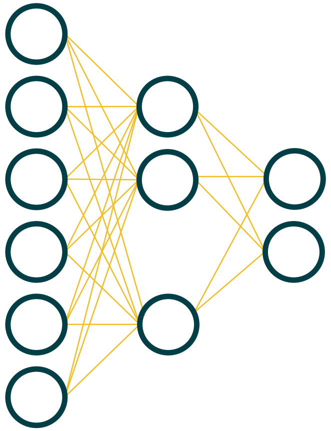
Model Pruning

2 STEP PROCESS

1. Reduce model size
2. Incrementally retrain model after pruning to recover accuracy

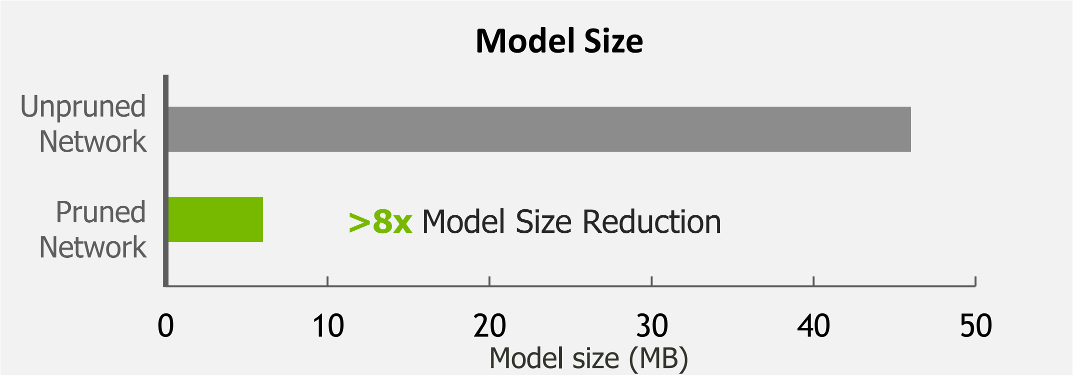


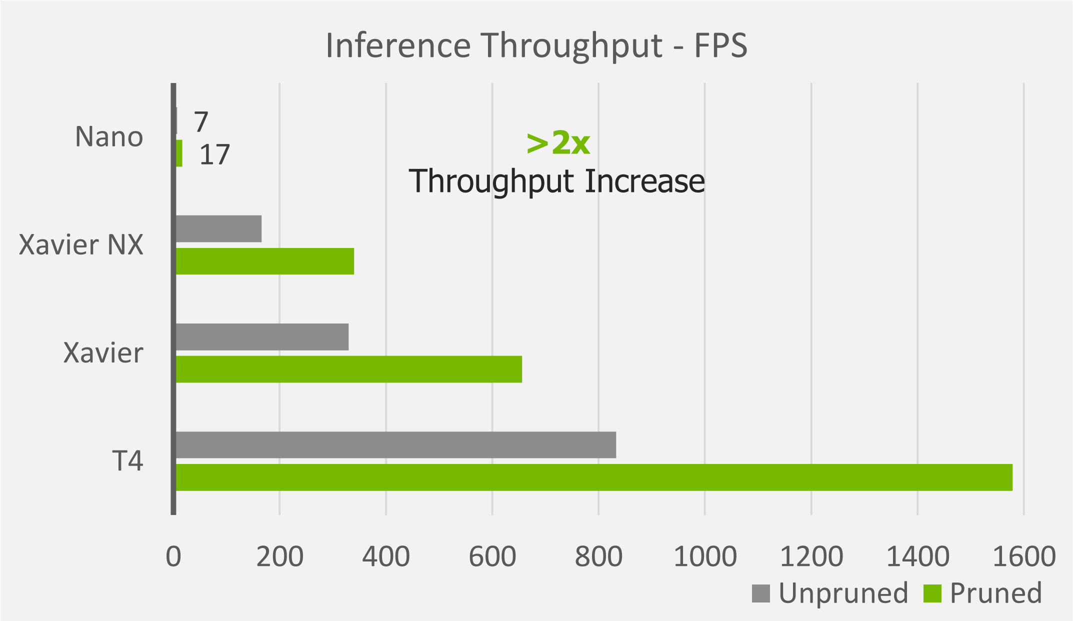
6 inputs, 6 neurons, 32 connections



6 inputs, 5 neurons, 24 connections

TrafficCamNet



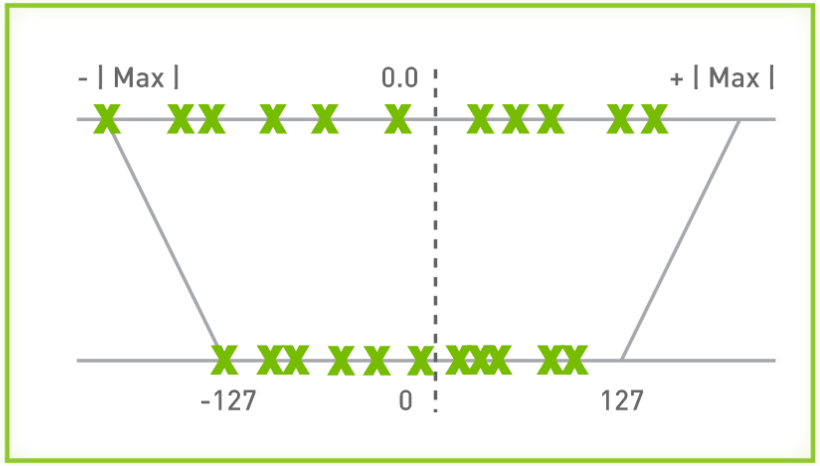


TrafficCamNet is DetectNet_v2 + ResNet18 based architecture (available on NGC)

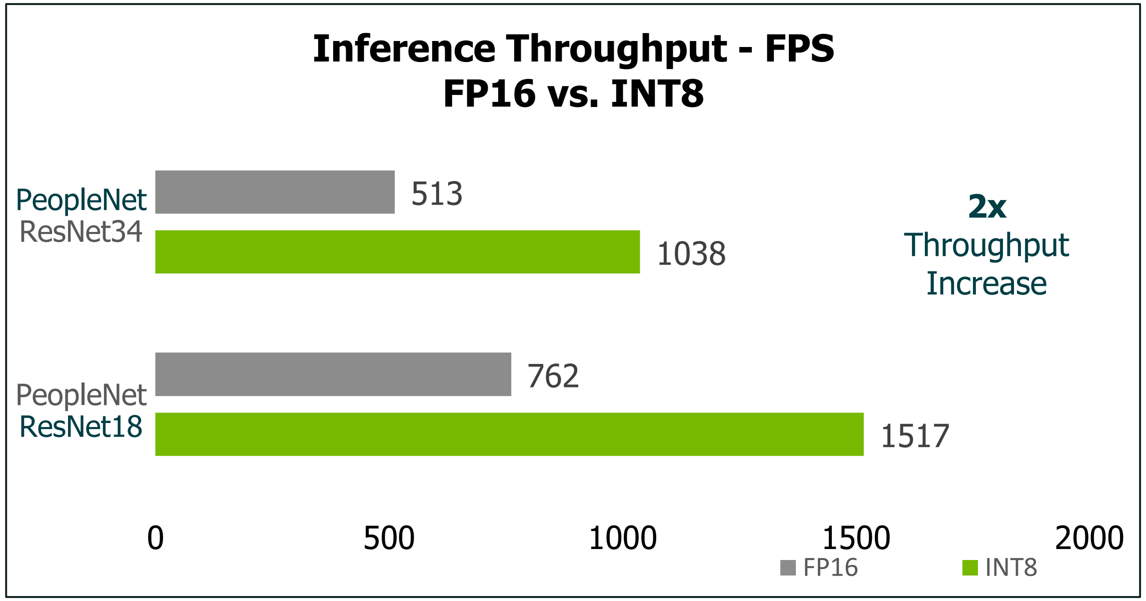
Quantization

Post Training Quantization (PTQ) for quantization after training is done

Quantization Aware Training (QAT) for quantization error from weights and tensors during training



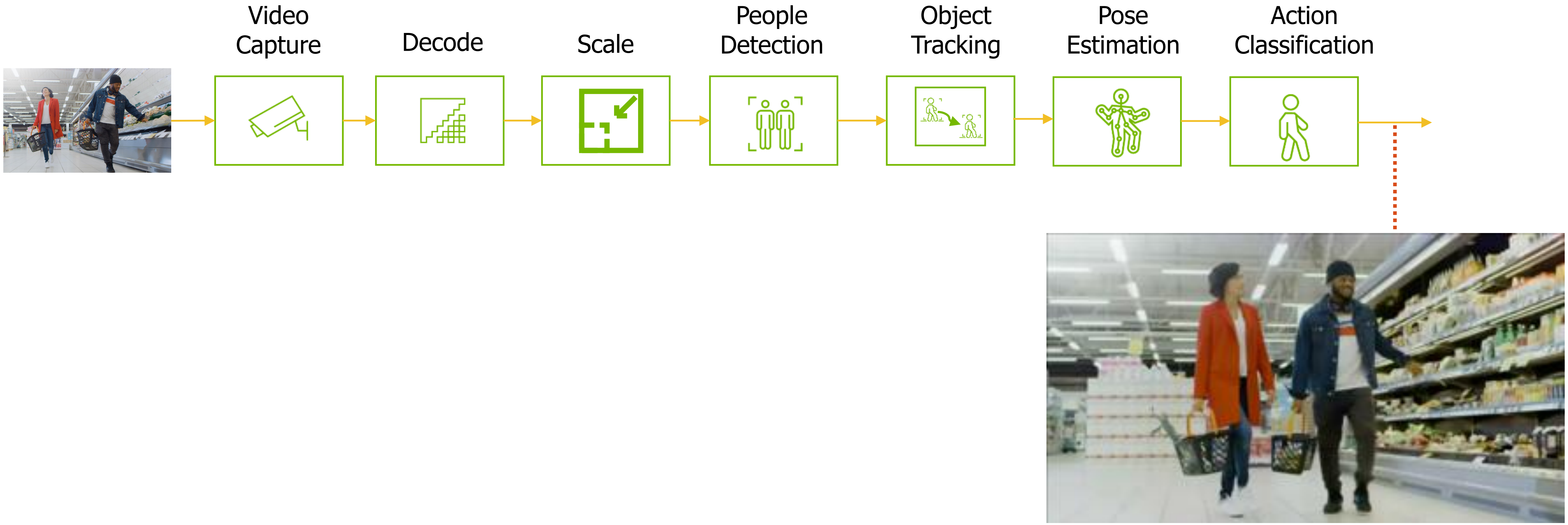
Transformation of floating-point weights to integer



<1% loss in accuracy between FP16 and INT8

https://developer.nvidia.com/blog/improving-int8-accuracy-using-quantization-aware-training-and-the-tao-toolkit/

Putting it all Together



Video Capture

Decode

Scale

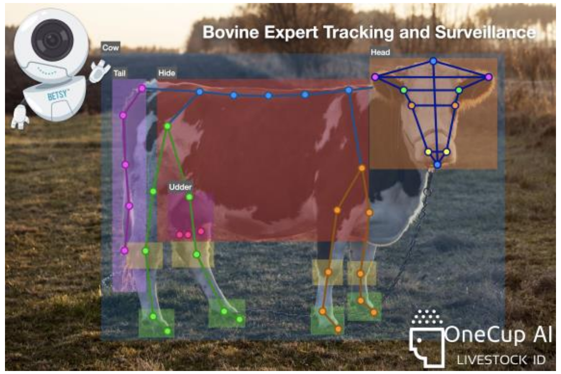
People Detection

Object Tracking

Pose Estimation

Action Classification

NVIDIA TAO – Accelerating AI Across Industries



One CupAI
Vision AI for Precision Agriculture

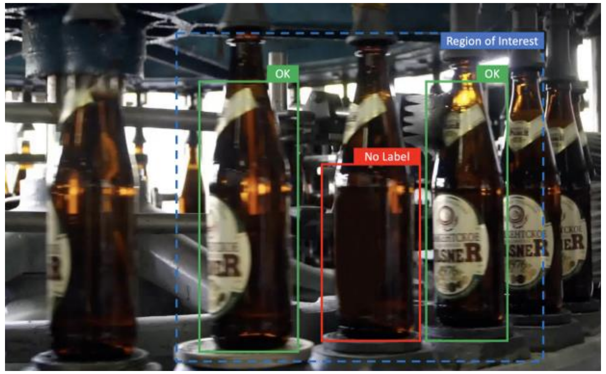
Remote tracking of animal health, growth and phenotype

Deployed an optimized solution that can run **inference on 1 TB** of data per day**, in just a few weeks**

6River Systems
Warehouse Logistics

Track objects in warehouse to optimize robot path planning and increase picking efficiency

Trained and deployed their model in application in just **weeks running with 30 parallel video streams instantaneously**



Mavenir
Quality Inspection

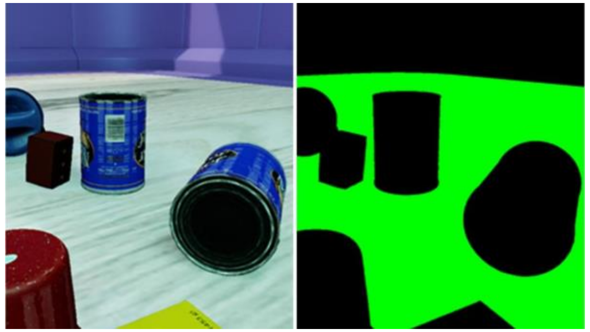
Detect defective bottles, labels before the final packaging step

Scaled across various use cases and **speed up development time by 3x**

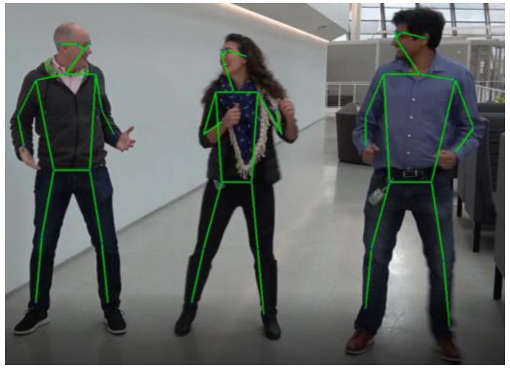
Summary

- The TAO Toolkit makes it easy for you to create custom, production-ready models for your speech and vision applications
  - Built on TensorFlow and PyTorch
- Diverse selection of pre-trained AI models
  - Removes the need for large training datasets
- Turnkey model optimization for inference
- Deploy easily with DeepStream for Vision AI applications at the edge

Resources

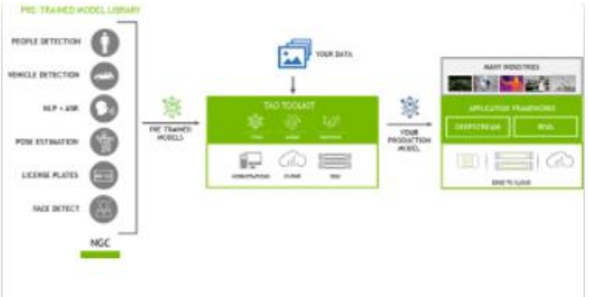


AI-Powered Robots with Synthetic Data



2D Pose Estimation Model with NVIDIA TAO Toolkit
Part 1 | Part 2

Train and Deploy Action Recognition Model



Supercharge your AI workflow with TAO Toolkit Whitepaper

https://developer.nvidia.com/tao-toolkit