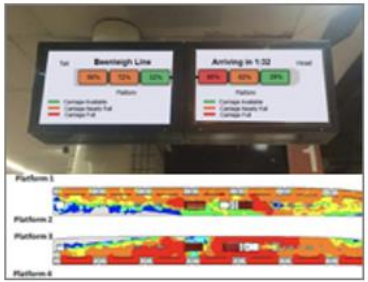
NVIDIA Transfer Learning Toolkit for Intelligent Video Analytics (IVA)

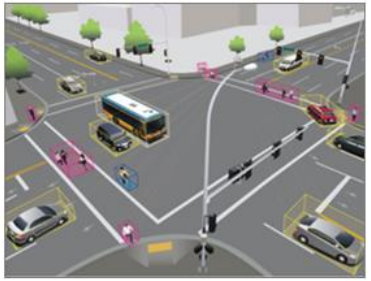
INTELLIGENT VIDEO ANALYTICS

Access Control

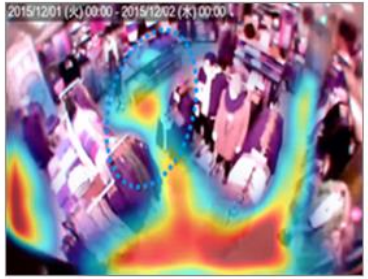


Public Transit

Parking Management



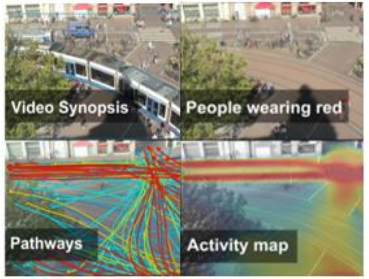
Traffic Engineering



Retail Analytics

Securing Critical Infrastructure

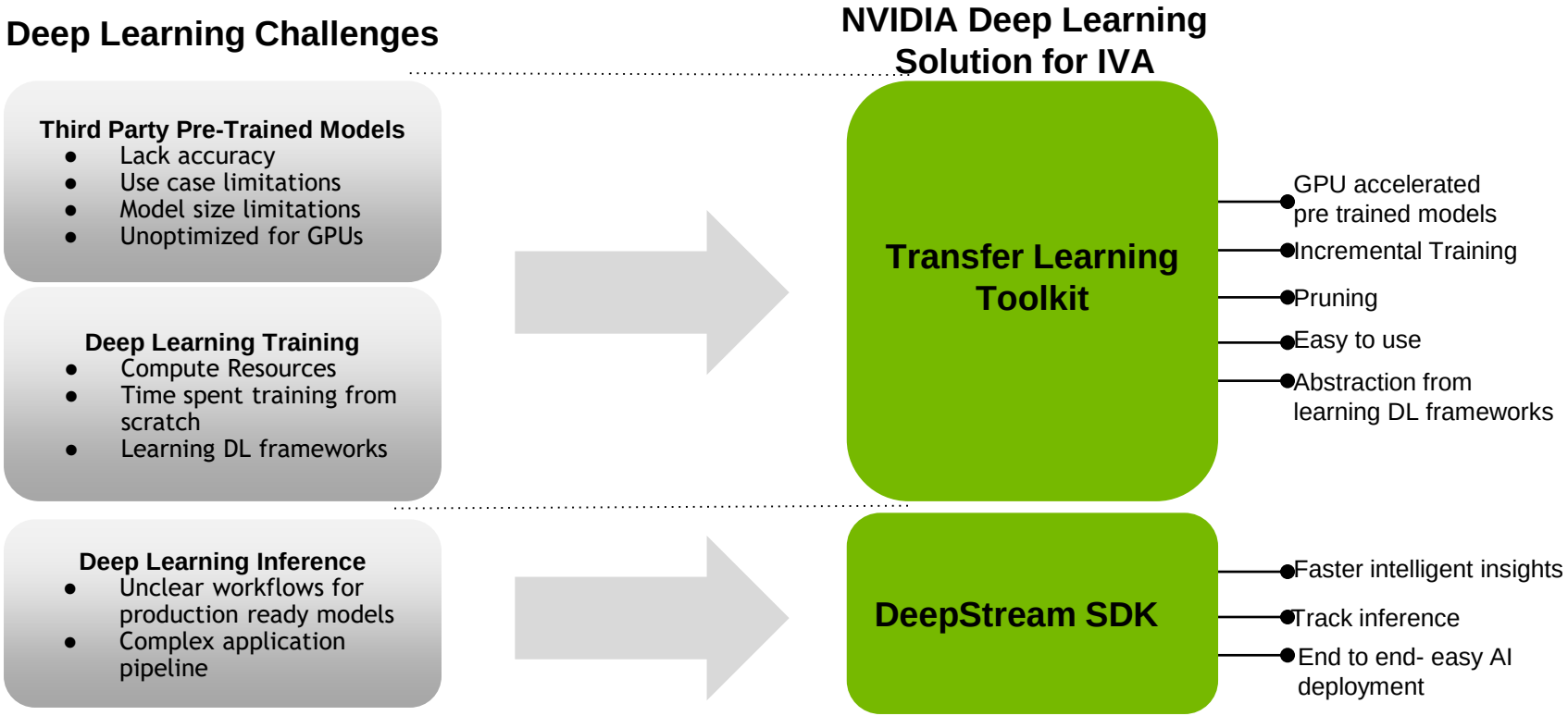
Managing Logistics



Forensic Analysis

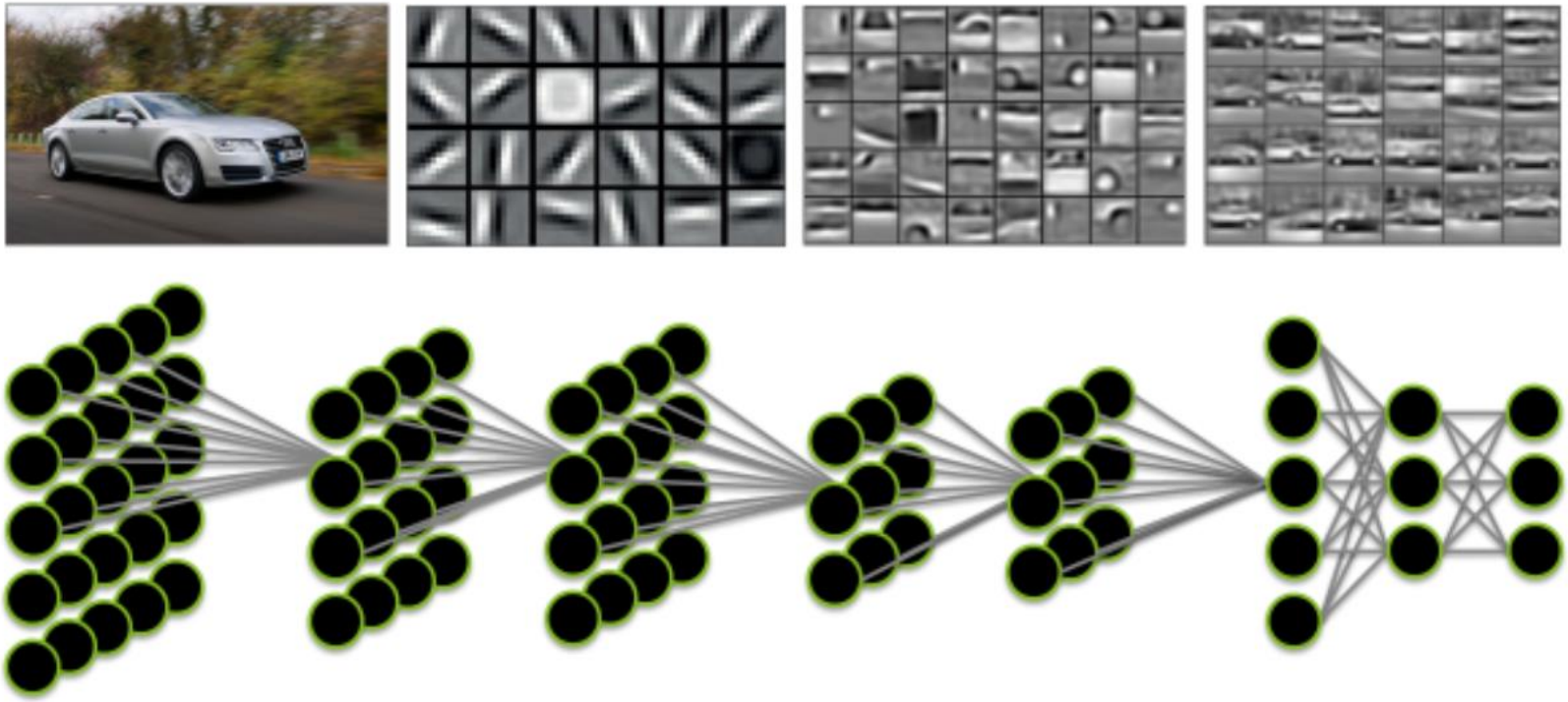
Increasing demand for deploying an efficient end to end deep learning workflow for IVA!

IVA Deep Learning Workflow Management



Accelerate deep learning training with pre-trained models and functions provided by Transfer Learning Toolkit

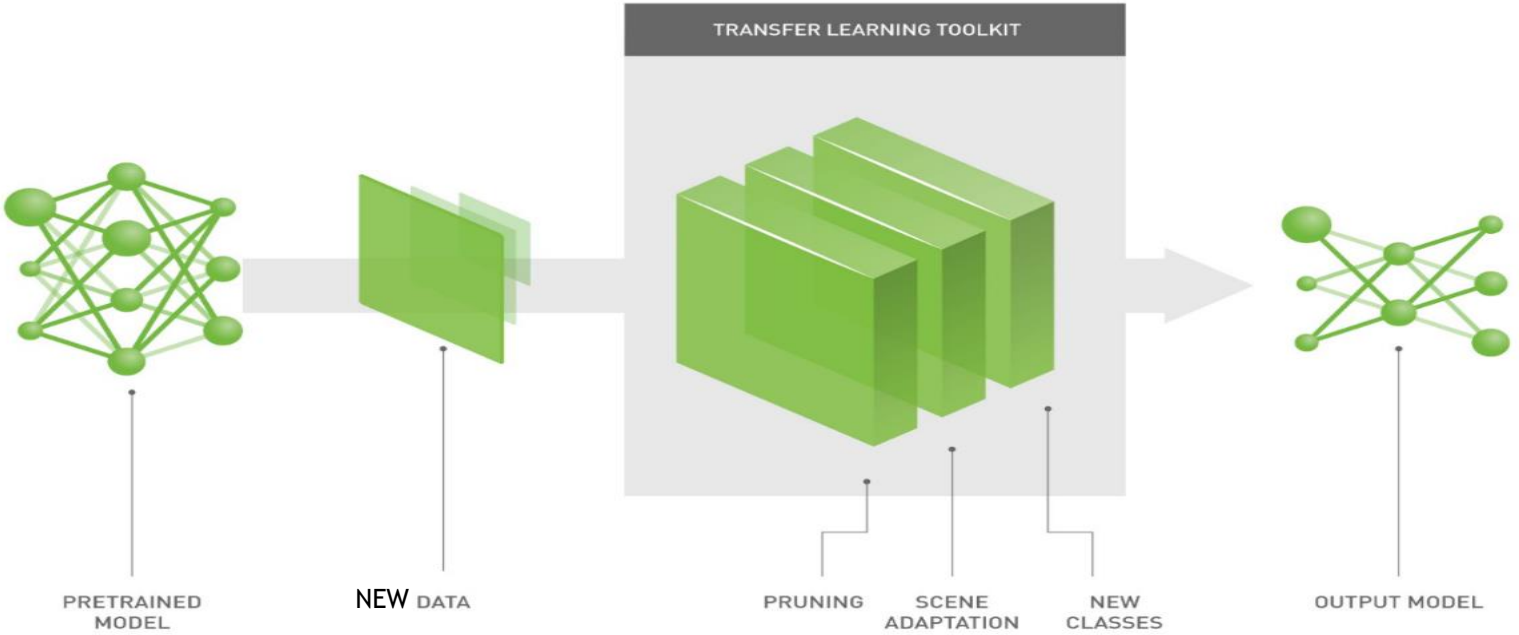
How Deep Learning Network Works



Transfer Learning is a process of transferring learned features from one model to another

Transfer Learning Toolkit

Fine Tuning * Pruning * Scene Adaptation * New Classes



Output Model ready to be deployed and integrated for us with DeepStream SDK 3.0 applications

USER PERSONAS IN DEEP LEARNING

Researchers

Goal: Develop algorithms

Data Scientists & Software Developers

Goal: Develop Applications for specific domain(IVA)

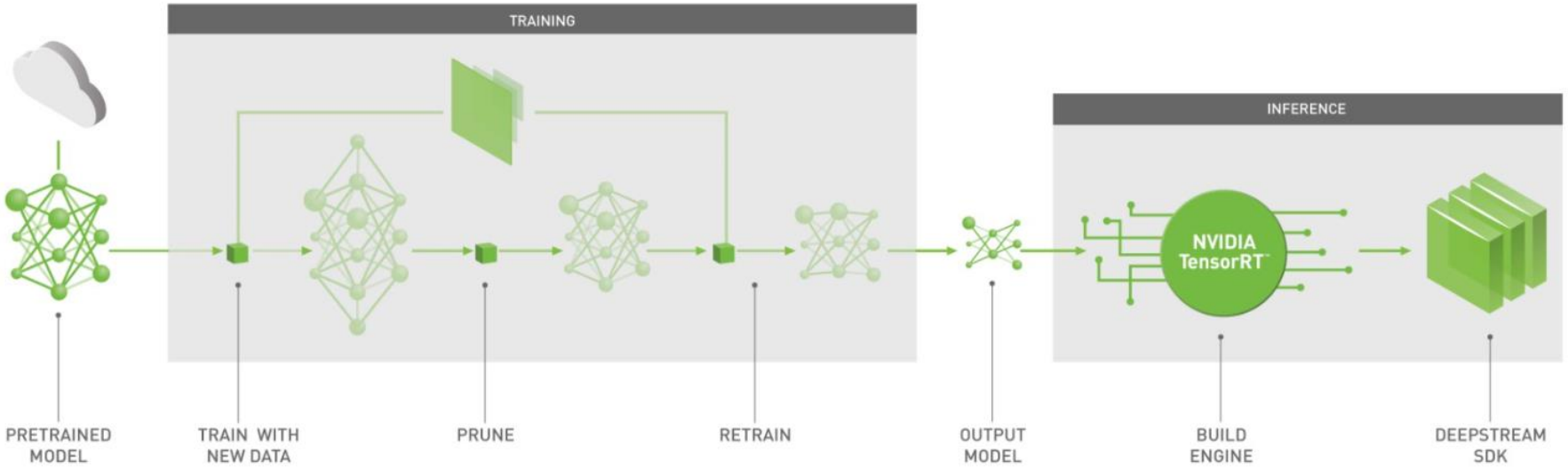
Sysadmins & IT DevOps

Goal: Manage infra for running compute jobs

Skills in Algorithms ----------> Skills in Domains & Applications----------> Skills in Systems

End to End NVIDIA Deep Learning Workflow

Pre-Trained model access from NGC * Training & adaptation * Applications ready to integrate with DeepStream



Accelerate time to market and save on compute resources!

NVIDIA TRANSFER LEARNING TOOLKIT FEATURES

Efficient Pre-trained Models

GPU-accelerated high performance models trained on large scale datasets.

Faster Inference with Model Pruning

Model pruning reduces size of the model resulting in faster inference

Training with Multiple GPUs

Re-training models, adding custom data for multi GPU training using an easy to use tool

Abstraction

Abstraction from having deep knowledge of frameworks, simple intuitive interface to the features

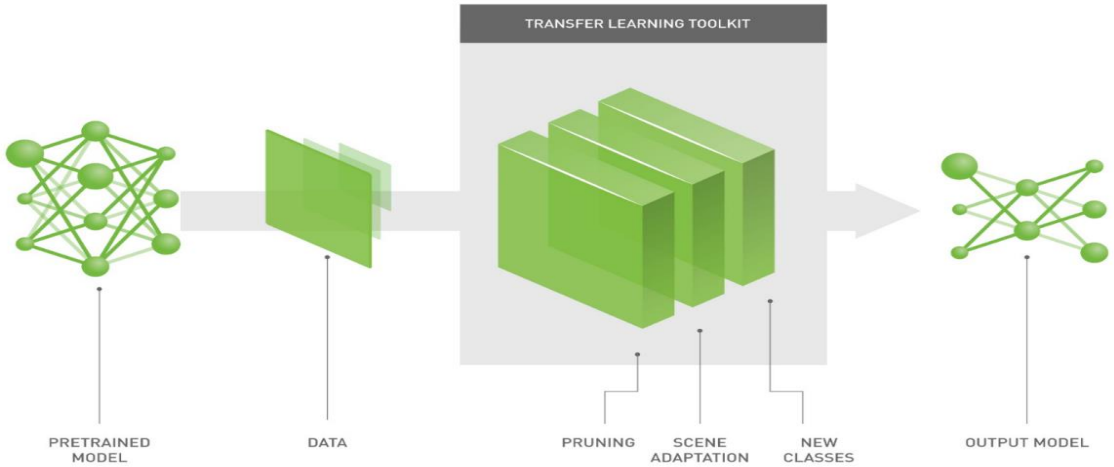
Containerization

Packaged in a container easily accessible from NVIDIA GPU Cloud website. All code dependencies are managed automatically

Integration

Models exported using TLT are easily consumable for inference with **Deep Stream SDK**

Now Generally Available! ngc.nvidia.com



Total 33 pretrained weights

Monochrome images

Better Metrics calculation

INT8 quantization

- Several object detection models
- INT8 quantization for high performance at deployment on Jetson
- Better Metrics support
- Monochrome images as input
- Support for Transfer learning freezing/unfreezing layers of the models
- Multi-GPU training support

Pretrained Models

All models are trained on google openimages public dataset
Available to download on ngc.nvidia.com

Image Classification

- ResNet10/18/50
- VGG16/19
- MobileNet V1/V2
- AlexNet
- SqueezeNet
- GoogLeNet

Object Detection

Faster RCNN supporting backbones:

- ResNet10/18/50
- VGG16/19
- GoogLeNet
- MobileNet V1/V2

DetectNet_v2 supporting backbones:

- ResNet10/18/50
- VGG 16/19
- GoogLeNet
- MobileNet V1/V2

SSD:

- ResNet10/18

Transfer Learning Toolkit

Key Capabilities

Pruning

Network compression

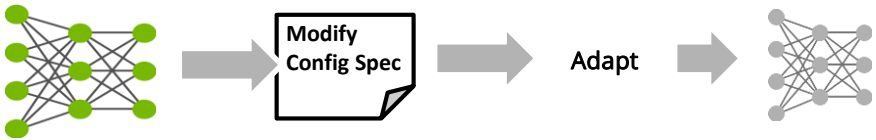
Scene Adaptation

Same network adapting to different angles and vantage points

Same network adapting to new data

Train with new data from another vantage point, camera location, or added attribute

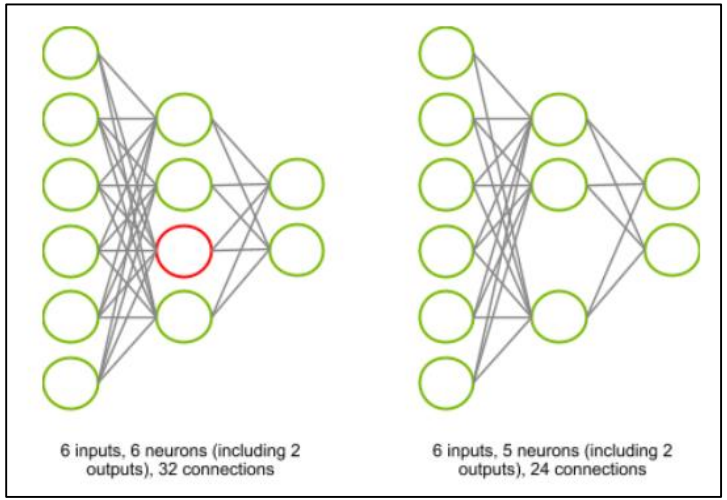
New Classes



Edited network detecting new class defined in config spec

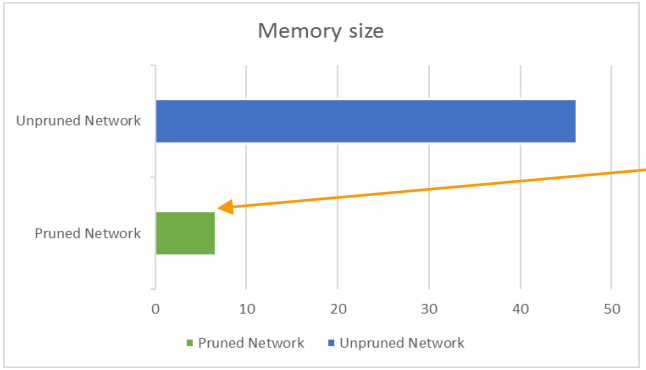
Pruning Models

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

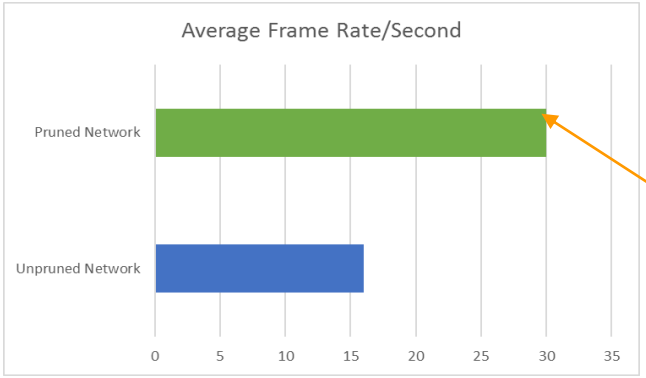


EXAMPLE

Network - ResNet 18 4-class (Car, Person, Bicycle, Road sign)
Memory size - 46.2 MB to 6.7 MB
FPS - 16 fps to 30 fps



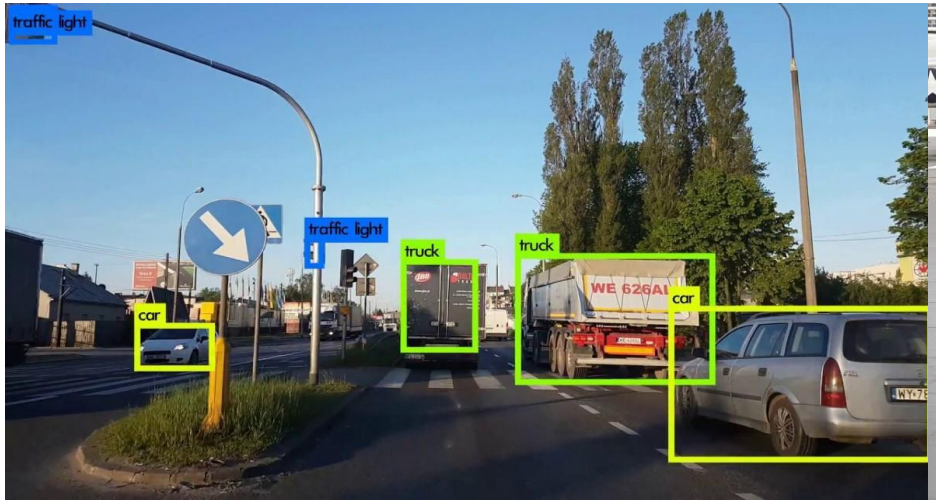
6.5x
Model Size Reduction



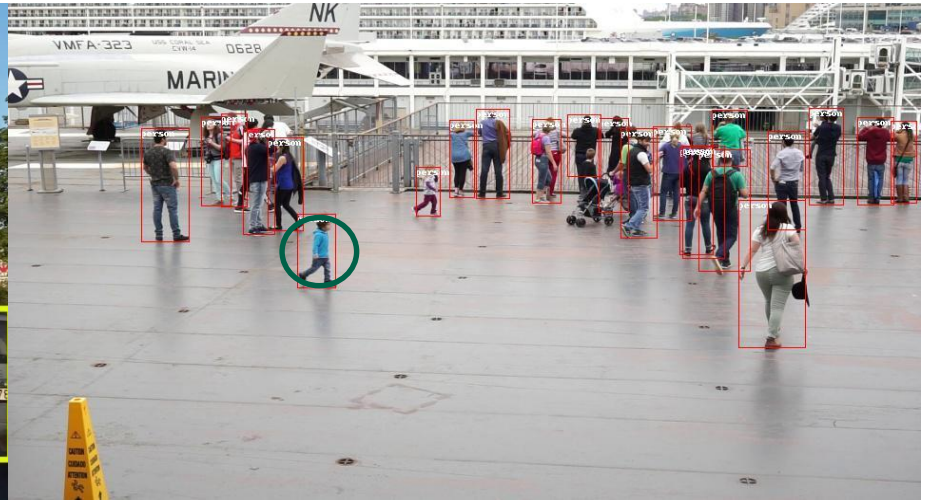
>2x
Throughput Increase

Scene Adaptation

Camera location vantage point



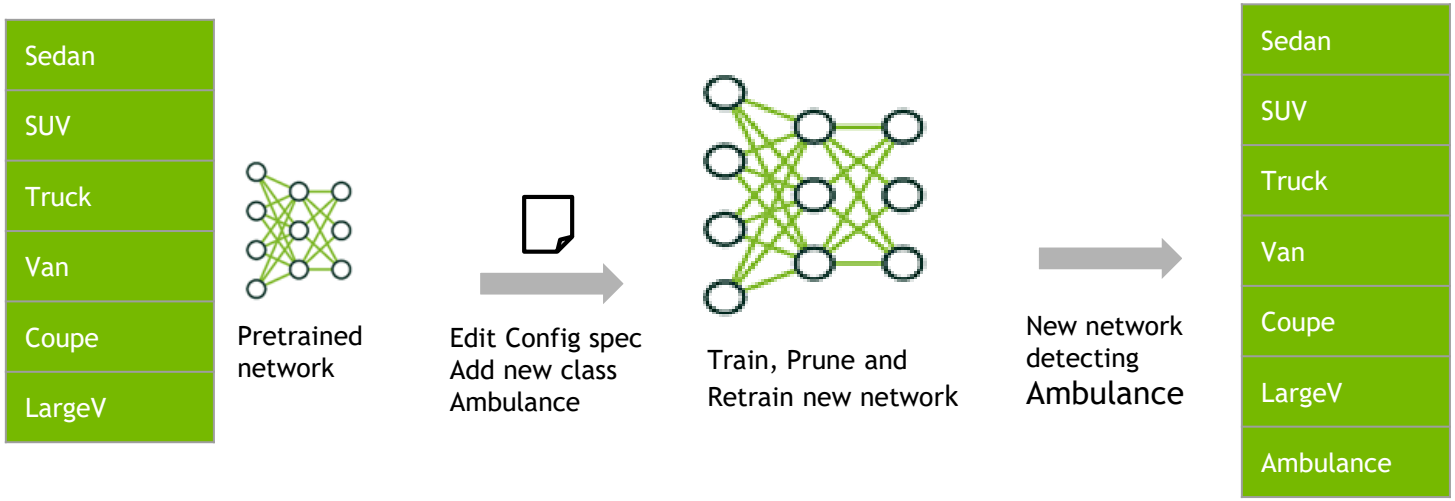
Person with blue shirt



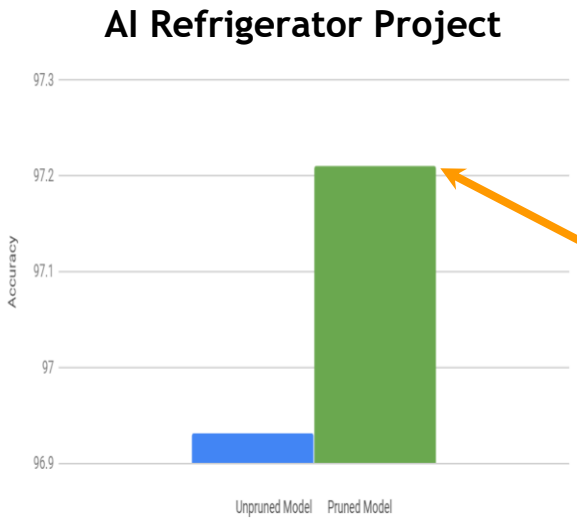
Train with new data from another vantage point, camera location, or added attribute

New Classes

Easy to modify models to add new classes



USE CASE 1: OBJECT DETECTION



54
Objects detected

5
Channels at 30 fps

Inference faster by 3x

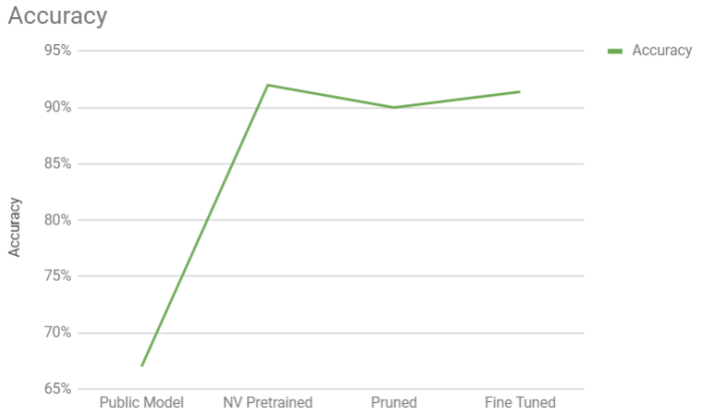
>90%
recall and precision

Memory footprint reduced by ~32x

QUICKLY PRUNE & FINE TUNE MODELS!

USE CASE 2: OBJECT CLASSIFICATION

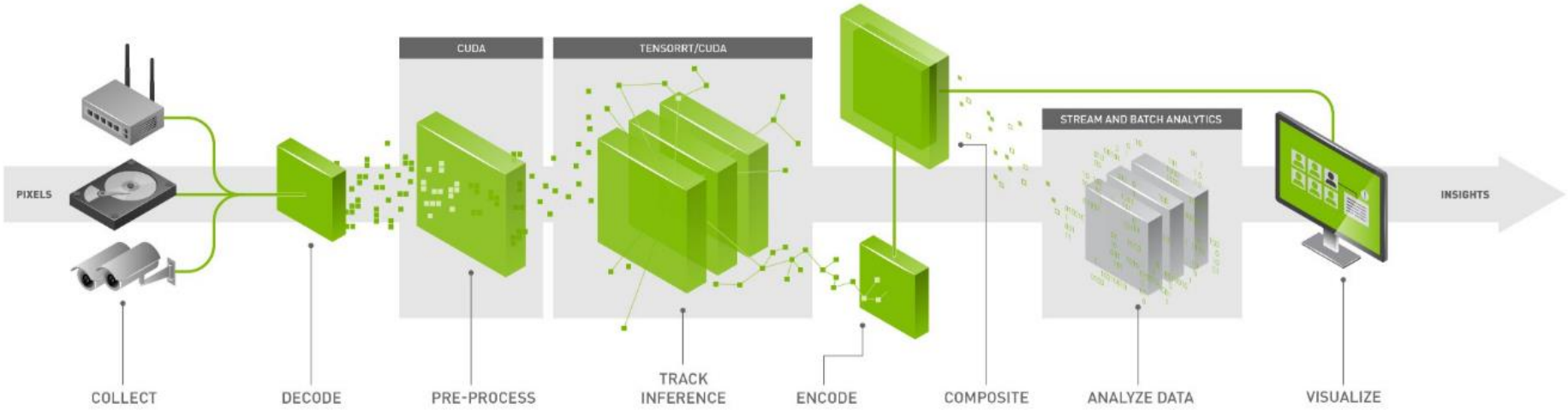
ResNet18 pre-trained model to determine object detected is a “real” person or not



Example Model performs ~5x better after pruning with toolkit

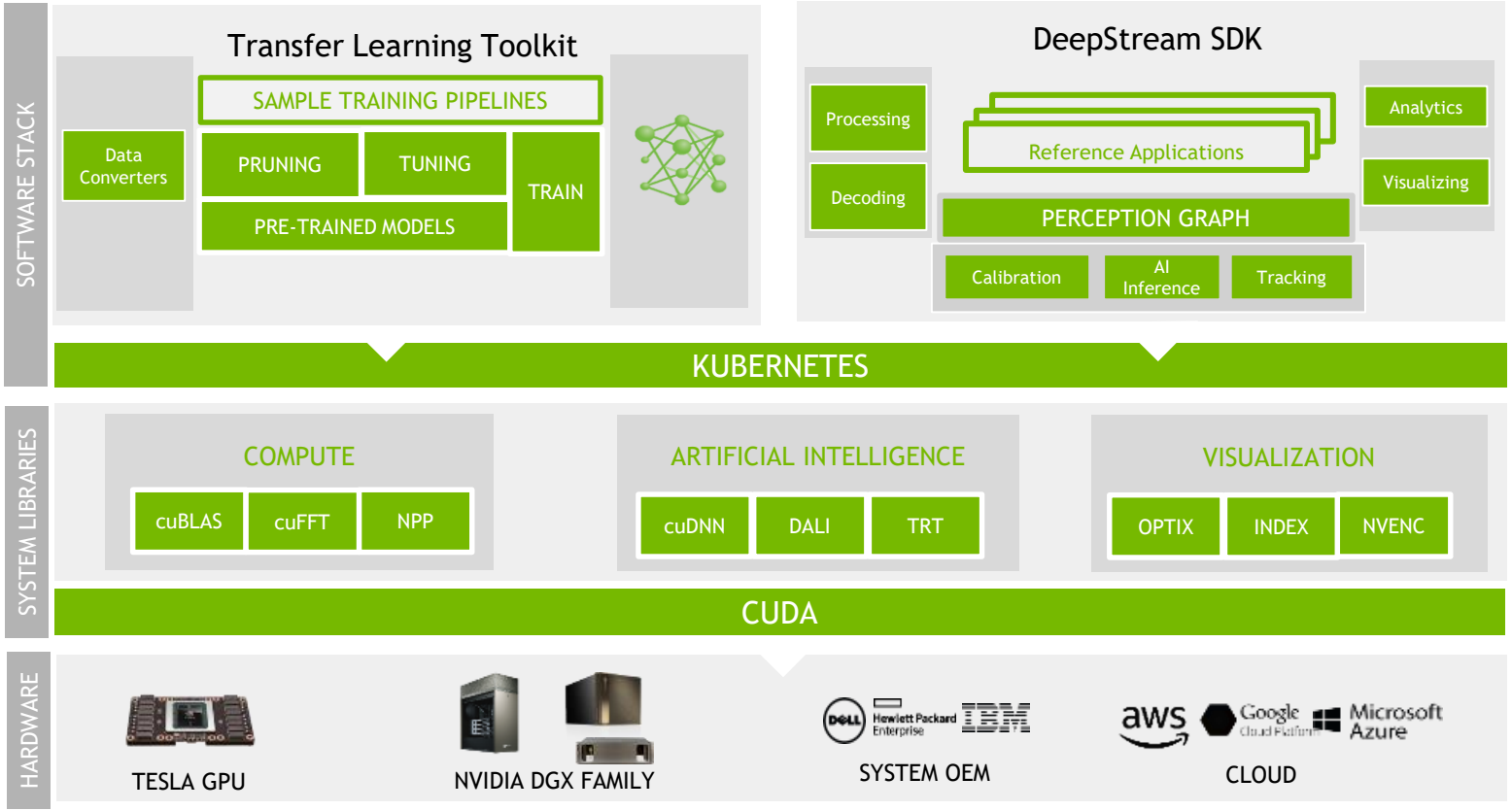

End to End Deep Learning Workflow for
Intelligent Video Analytics

Deep Learning for IVA
Transfer Learning Toolkit With DeepStream SDK



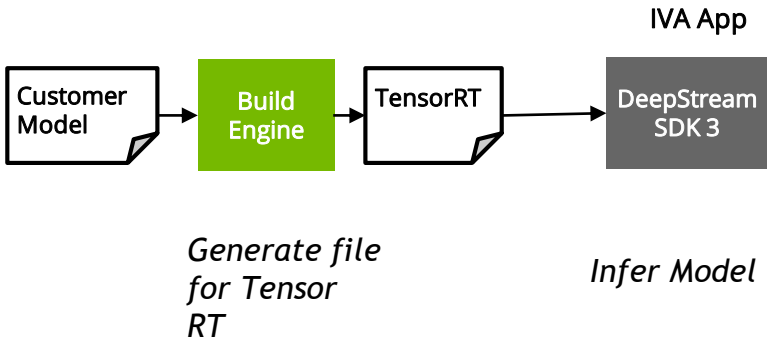
Accelerate building and deploying heterogeneous applications for IVA use cases with TLT & DeepStream SDK

Deep Learning for IVA



Inference with NVIDIA DeepStream SDK

- NVIDIA DeepStream SDK is a solution for designing applications for Intelligent Video Understanding
- Transfer Learning Toolkit provides API to easily integrate custom model for Inference via DeepStream
- Start with the NVIDIA Pre-trained models
- Use Transfer learning Toolkit to adapt to custom data, prune, retrain and export
- Use plugin-ins in DeepStream SDK to quickly deploy applications for inference



Getting Started: Transfer Learning Toolkit

- Download Transfer Learning Toolkit today!
- Deploy end to end IVA solution with NVIDIA DeepStream. Download DeepStream
- Documentation available online
- Blogs:
  - Tutorial
  - What is Transfer Learning?
  - Pruning Models with Transfer Learning Toolkit