ACCELERATE DEEP LEARNING INFERENCE USING INTEL TECHNOLOGIES

OPTIMIZATION: TOOLS AND TECHNIQUES

February 2020

OPTIMIZATION NOTICE

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness or any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice. Notice Revision #20110804.

LEGAL NOTICES AND DISCLAIMERS (1 OF 2)

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software, or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at www.intel.com.

Performance estimates were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

This document contains information on products, services, and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications, and roadmaps.

Any forecasts of goods and services needed for Intel's operations are provided for discussion purposes only. Intel will have no liability to make any purchase in connection with forecasts published in this document.

Arduino*101 and the Arduino* infinity logo are trademarks or registered trademarks of Arduino, LLC.

Altera, Arria, the Arria logo, Intel, the Intel logo, Intel Atom, Intel Core, Intel Nervana, Intel Xeon Phi, Movidius, Saffron, and Xeon are trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

*Other names and brands may be claimed as the property of others.

LEGAL NOTICES AND DISCLAIMERS (2 OF 2)

This document contains information on products, services, and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications, and roadmaps. Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software, or service activation. Learn more at intel.com or from the OEM or retailer. No computer system can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit www.intel.com/performance.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

Statements in this document that refer to Intel's plans and expectations for the quarter, the year, and the future, are forward-looking statements that involve a number of risks and uncertainties. A detailed discussion of the factors that could affect Intel's results and plans is included in Intel's SEC filings, including the annual report on Form 10-K.

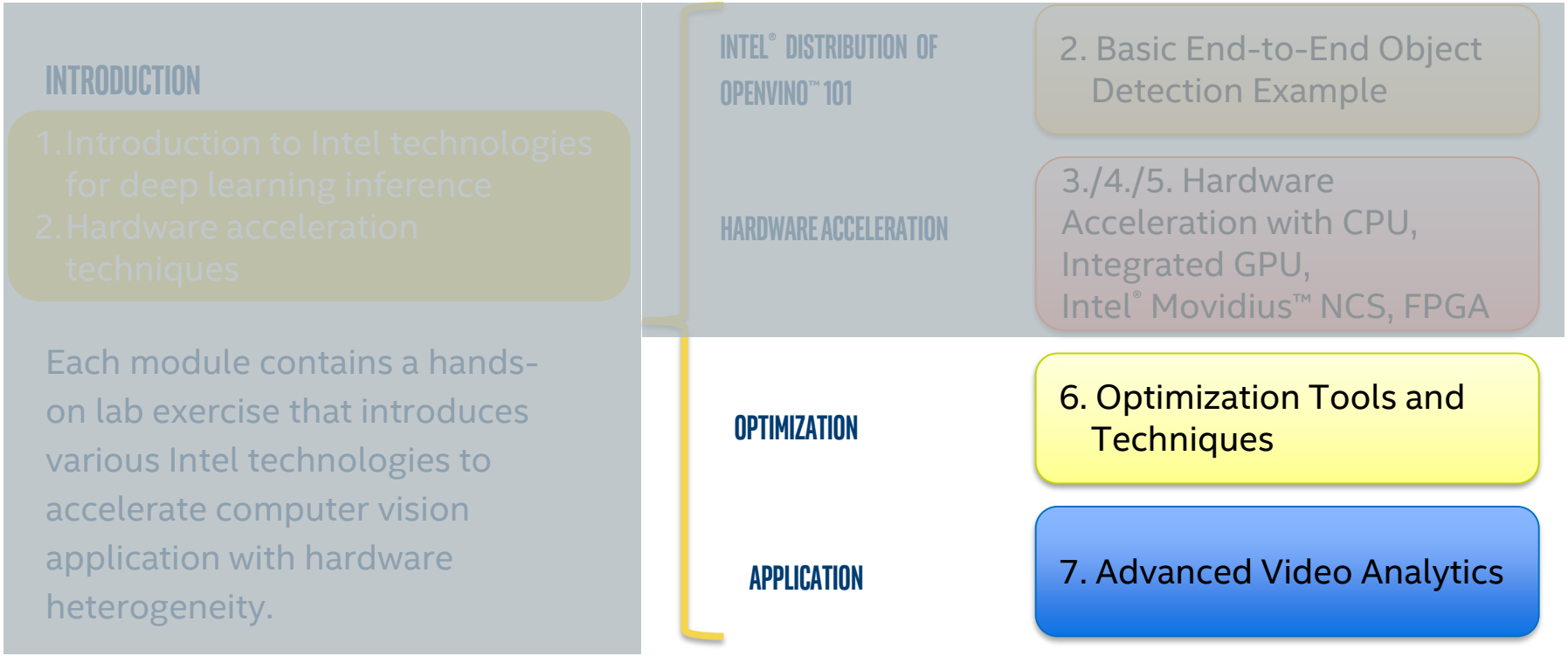
The products described may contain design defects or errors, known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request. Performance estimates were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document. Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

Intel, the Intel logo, Pentium, Celeron, Atom, Core, Xeon, Movidius, Saffron, and others are trademarks of Intel Corporation in the United States and other countries.

*Other names and brands may be claimed as the property of others.

SMART VIDEO WORKSHOP OVERVIEW



AGENDA

- Pick the Right Model
- Use DL Workbench
- Don’t Infer If Not Needed
- Use Command Line Deployment Manager

PICK THE RIGHT MODEL

USE/TRAIN A MODEL WITH THE RIGHT PERFORMANCE PLUS ACCURACY TRADEOFFS.

Performance is based on many factors:

- Topography complexity/layer implementation plus scheduling
- Number of color channels (that is, BGR vs. grayscale)
- Model resolution

EXERCISE: RANGE OF MODEL PERFORMANCE

FOCUS ON THE INFERENCE TIMING

	CPU ms/frame	GPU ms/frame	Intel® Movidius™ Myriad™ X ms/frame
ssd512			
ssd300			
Mobilnet-ssd*			

EXERCISE: RANGE OF MODEL PERFORMANCE

FOCUS ON THE INFERENCE TIMING

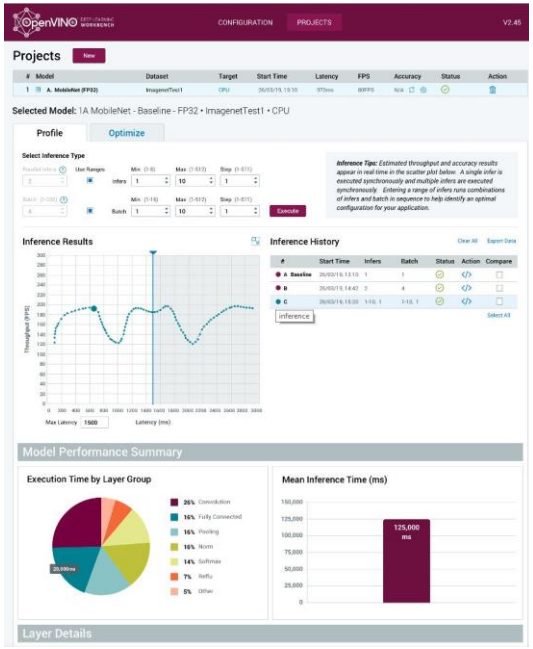
	CPU ms/frame	GPU ms/frame	Intel® Movidius™ Myriad™ X ms/frame
ssd512	1260.35 ms/frame	649.604 ms/frame	1385.13 ms/frame
ssd300	404.721 ms/frame	227.864 ms/frame	608.919 ms/frame
Mobilnet-ssd*	18.8134 ms/frame	20.8313 ms/frame	38.5964 ms/frame

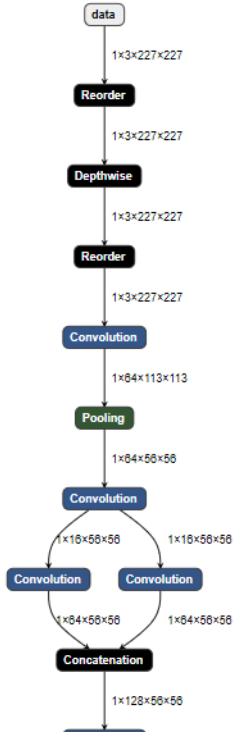
USE DL WORKBENCH

DEEP LEARNING WORKBENCH

Deep Learning Workbench capabilities

- Web-based tool - UI extension of Intel® Distribution of OpenVINO™ toolkit functionality
- Visualizes performance data for topologies/ layers to aid in model analysis
- Automate analysis for optimal performance configuration (streams, batches, latency)
- Experiment with int8 calibration for optimal tuning
- Provide accuracy info through accuracy checker
- Direct access to Models from public set of Open Model Zoo





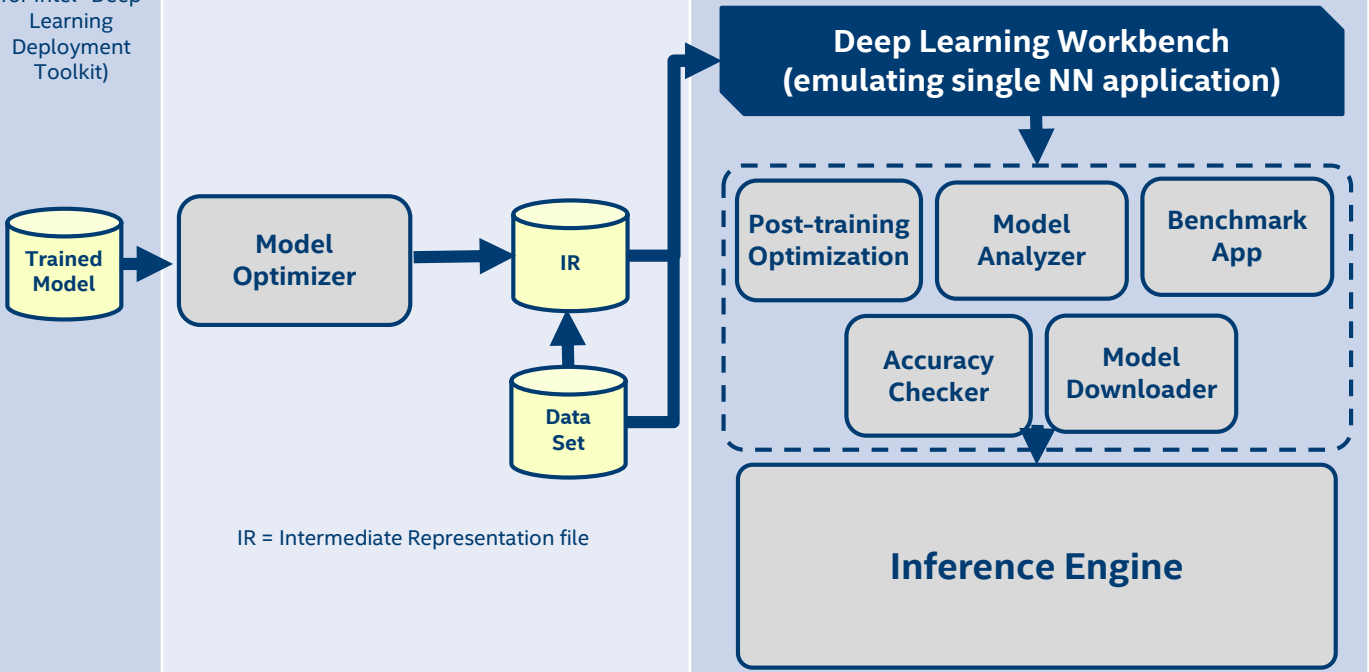
DEEP LEARNING WORKBENCH DATA FLOW

STEP 0:
Train model with FW (Out-of-scope for Intel® Deep Learning Deployment Toolkit)

STEP 1:
Convert model to IR file with use of Model Optimizer

STEP 2:
Import Intermediate Representation (IR) file & dataset - then conduct multiple performance experiments

STEP 3:
Deploy application



IR = Intermediate Representation file

WORKBENCH INTERFACES WITH KEY COMPONENTS

- **Post-Training Optimization Toolkit** – Convert a model into a more hardware-friendly representation by applying specific methods that do not require re-training, for example, post-training quantization.
- **Model Analyzer** – Provides theoretical data on models: computational complexity (flops), number of neurons, memory consumption.
- **Benchmark App** – Helps measure performance (throughput, latency) of a model, get performance metrics per layer and overall basis.
- **Accuracy Checker Tool** – Check for accuracy of the model (original and after conversion) to IR file using a known data set.
- **Model Downloader** – Provides an easy way of accessing a number of public neural network models as well as a set of pre-trained Intel models

Installation & Distribution

Intel® Distribution of OpenVINO™ toolkit

- Build your local docker image from package (build scripts)
- Build your local docker image by copying and running dockerfile from documentation

Download docker image from DockerHub

- Work in progress – will depend on decision of OpenSource PDT
- Can be used for internal experiments (at least)

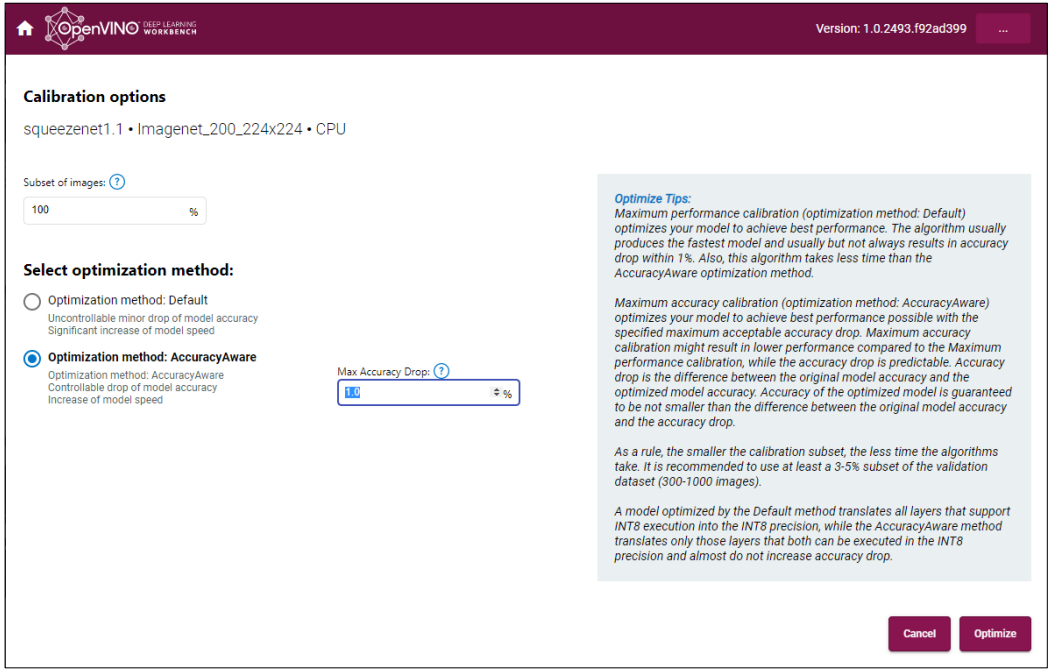
IR = Intermediate Representation file

CONVERT MODEL TO INT8 USING 2 NEW CALIBRATION ALGORITHMS

Audience: New and experienced users of OpenVINO and DL Workbench

Problem: *Provide an easier to use (non-command line) interface to new Post training optimization (calibration) tool and calibration algorithms*

UseCase: *OpenVINO user can convert her model to Int8 using 2 new algorithms achieving this goal using purely UI.*

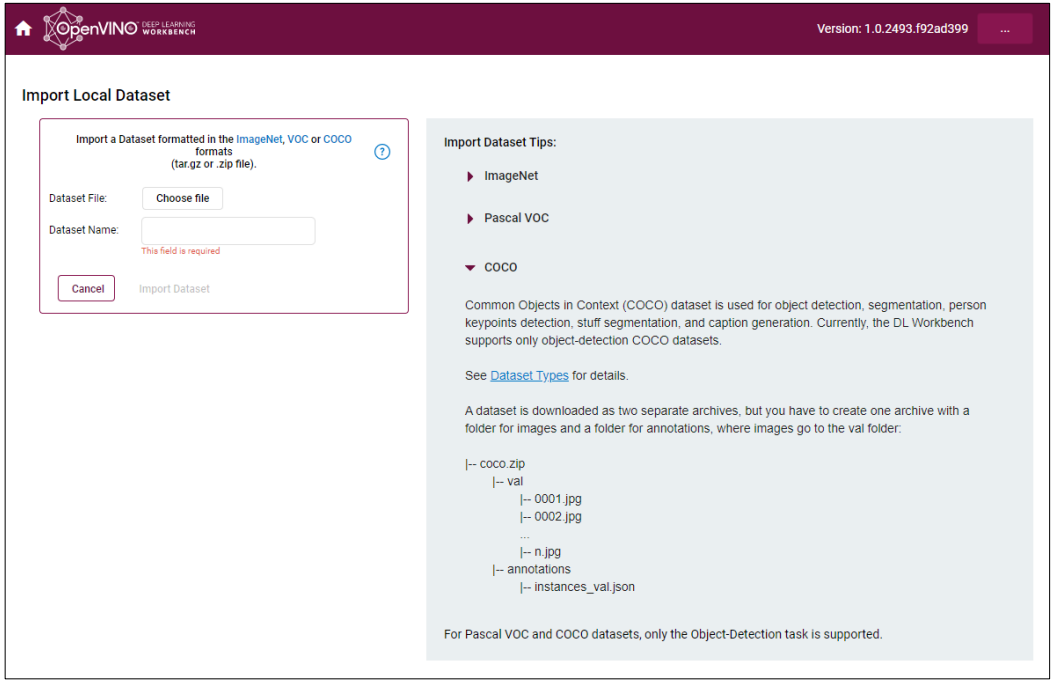


IMPORT DATASET IN COCO FORMAT TO USE WITH MODEL

Audience: New and experienced users of OpenVINO and DL Workbench

Problem: Currently DL Workbench supports only ImageNet format dataset(s) or Pascal VOC format dataset(s) in flow. COCO dataset extends the list of supported formats giving additional freedom in selecting dataset for experiments.

UseCase: *Applicable for all use cases provided by DL Workbench.*

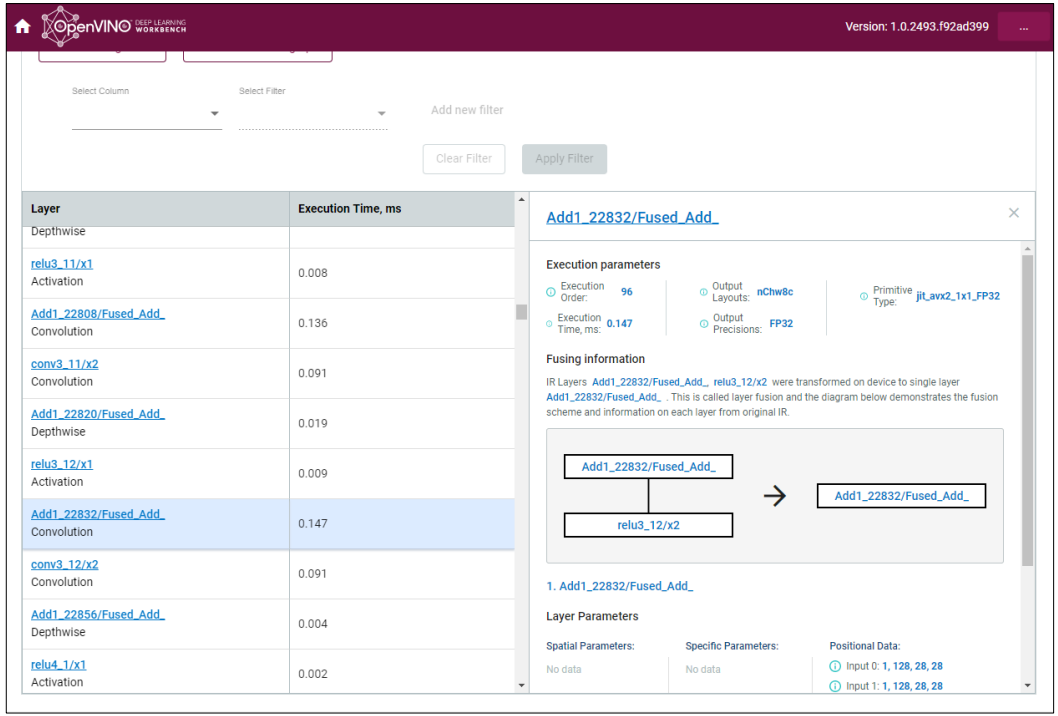


IMPROVED PER-LAYER DATA VISUALIZATION AND COMPARISON MODE. MULTIPLE UX IMPROVEMENTS.

Audience: New and existing users of DL Workbench

Problem: *Address the number of UX items. Improve visualization of table with per-layer information*

UseCase: *User have more intuitive visualization of the per-layer information including fuzing of layers on target HW and comparison mode for the models.*



DON’T INFER IF NOT NEEDED

DETERMINE IF THERE IS NOTHING TO SEE

Inference is expensive to run each frame. It can save time to not run when there is nothing to identify.

- Check motion vectors
- Frame sizes
- bgsubmog
- SAD

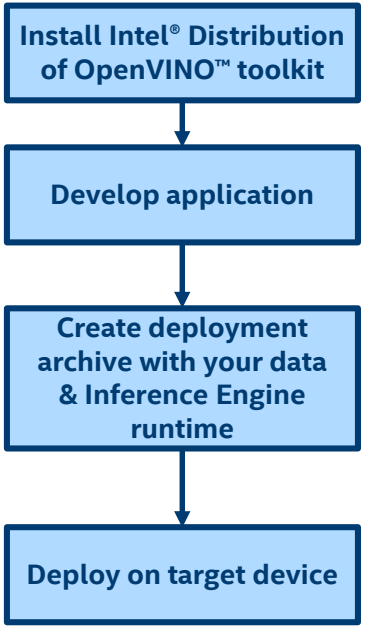
These methods can be several orders of magnitude less expensive than inference. Use techniques to increase the total # of streams a system can watch.

USE COMMAND LINE DEPLOYMENT MANAGER

COMMAND LINE DEPLOYMENT MANAGER

- Generate an optimal, minimized runtime package for selected target device.
- Deploy Inference Engine with pre-compiled application-specific data such as models, config, and a subset of required hardware plugins.
- Achieve deployment footprint to be several times smaller than the development footprint.

For more details, see Introduction to CLI Deployment Manager



Target	Size, MB
CPU only	65
GPU only	26
Myriad only	22
HDDL only	27
GNA only	15

Measurements for deployment archives based on 2019 R3

LAB5 - OPTIMIZING COMPUTER VISION APPLICATIONS

URL: https://github.com/intel-iot-devkit/smart-video-workshop/blob/master/optimization-tools-and-techniques/README.md

Objective: This tutorial shows some techniques to get better performance for computer vision applications with the Intel® Distribution of OpenVINO™ toolkit.

Estimated Complete Time: 40min

ADVANCED VIDEO ANALYTICS

SECURITY BARRIER DEMO

February 2020

VIDEO ANALYTICS IN INTEL® DISTRIBUTION OF OPENVINO™ TOOLKIT

Topology	Type	Description
license-plate-recognition-barrier-0001	ocr	Chinese license plate recognition.
vehicle-attributes-recognition-barrier-0010	object_attributes	Vehicle attributes recognition with modified RESNET10* backbone.
vehicle-license-plate-detection-barrier-0007	detection	Multiclass (vehicle, license plates) detector based on RESNET10 plus SSD.

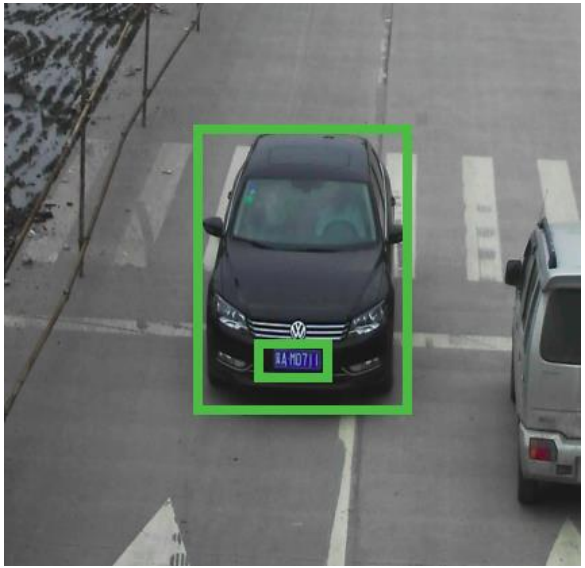
VEHICLE-ATTRIBUTES-RECOGNITION-BARRIER-0010
USE CASE/HIGH-LEVEL DESCRIPTION

Vehicle attributes classification algorithm for a traffic analysis scenario.

Type: regular
Color: black

VEHICLE-LICENSE-PLATE-DETECTION-BARRIER-007
USE CASE/HIGH-LEVEL DESCRIPTION

RESNET* 10 plus SSD-based vehicle and (Chinese) license plate detector for "Barrier" use case.



LICENSE-PLATE-RECOGNITION-BARRIER-0001
USE CASE/HIGH-LEVEL DESCRIPTION

Small-footprint network trained E2E to recognize Chinese license plates in traffic scenarios.

Note: The license plates in the image are modified from the originals.

SECURITY BARRIER DEMO

LAB7 - ADVANCED VIDEO ANALYTICS

URL: https://github.com/intel-iot-devkit/smart-video-workshop/blob/master/advanced-video-analytics/multiple_models.md

Objective: The tutorial shows some techniques for developing advanced video analytics applications.

Estimated Complete Time: 20min


intel