ACCELERATE AI INFERENCE FROM CLOUD TO EDGE

WITH ONNX* RUNTIME + OPENVINO™ TOOLKIT

DATE: JUNE, 2020

NOTICES AND DISCLAIMER

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

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.

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.

ONNX* EXCHANGE

ONNX

What it is

ONNX is an open ecosystem that empowers AI developers to choose the right tools as their project evolves. ONNX provides an open source format for AI models, both deep learning and traditional ML. It defines an extensible computation graph model, as well as definitions of built-in operators and standard data types. ONNX is currently focused on the capabilities needed for inferencing (scoring).

Target audience

- Computer vision, machine learning and deep learning software developers
- Data scientists
- OEMs, ISVs, System Integrators

Usages

AI for robotics, retail, healthcare, security surveillance, office automation, transportation, non-vision use cases (speech, NLP, Audio, text) & more.

AI FRAMEWORK INTEROPERABILITY – COMMON FORMAT

TOOLS TO CONVERT MODEL FORMATS TO ONNX

MODEL CATALOG THROUGH ONNX MODEL ZOO

STREAMLINING PATH FROM PROTOTYPE TO PRODUCTION

Homepage ▸ onnx.ai

Github ▸ github.com/onnx/onnx

ONNX Model Zoo Github ▸ https://github.com/onnx/models

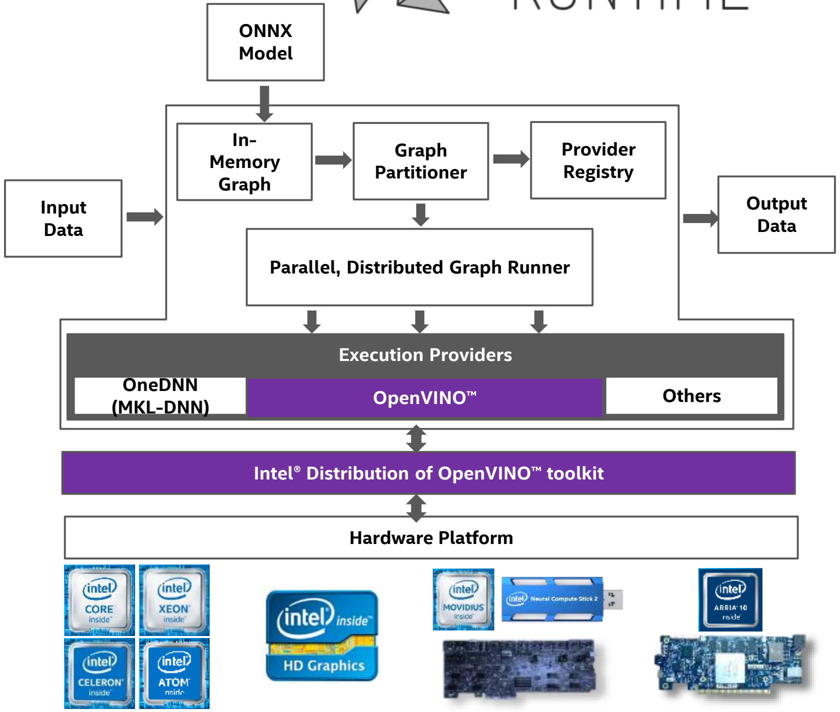
ONNX* RUNTIME

What it is

ONNX Runtime is a performance-focused complete scoring engine for Open Neural Network Exchange (ONNX) models, with an open extensible architecture to continually address the latest developments in AI and Deep Learning. ONNX Runtime stays up to date with the ONNX standard and supports all operators from the ONNX v1.2+ spec with both forwards and backwards compatibility. Execution Provider plugin allows the support of ONNX RT for Intel® Distribution of OpenVINO™ toolkit.

- ✓ BUILT SPECIFICALLY FOR ONNX FORMAT MODELS
- SUPPORTS EXECUTION ON MANY TYPES OF HARDWARE
- COMPLETELY OPEN SOURCED ON GITHUB

ONNX RUNTIME



Homepage ▸ onnx.ai

Github ▸ github.com/Microsoft/onnxruntime

INTEL® DISTRIBUTION OF OPENVINO™ TOOLKIT

What it is

A toolkit to accelerate development of **high performance computer vision** & **deep learning inference into vision/AI applications** used from edge to cloud. It enables deep learning on hardware accelerators and easy deployment across multiple types of Intel® platforms.

Target audience

- Computer vision, deep learning software developers
- Data scientists
- OEMs, ISVs, System Integrators

Usages

AI for robotics, retail, healthcare, security surveillance, office automation, transportation, non-vision use cases (speech, NLP, Audio, text) & more.

OpenVINO™

HIGH PERFORMANCE, PERFORM AI AT THE EDGE

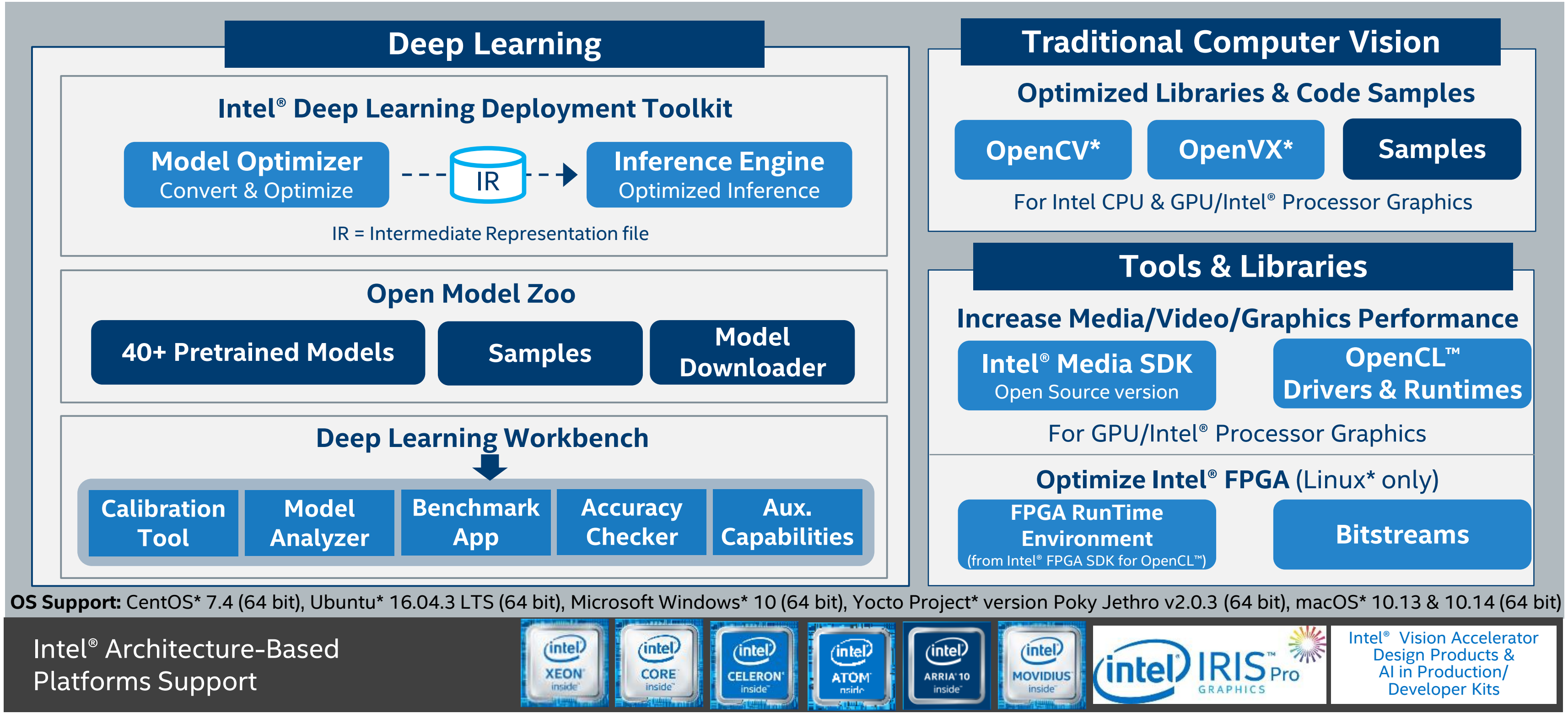
STREAMLINED & OPTIMIZED DEEP LEARNING INFERENCE

HETEROGENEOUS, CROSS-PLATFORM FLEXIBILITY

Free Download ▸ software.intel.com/openvino-toolkit

Open Source version ▸ 01.org/openvinotoolkit

INTEL® DISTRIBUTION OF OPENVINO™ TOOLKIT



Deep Learning

Intel® Deep Learning Deployment Toolkit

Model Optimizer
Convert & Optimize

IR

Inference Engine
Optimized Inference

IR = Intermediate Representation file

Open Model Zoo

40+ Pretrained Models

Samples

Model Downloader

Deep Learning Workbench

Calibration Tool

Model Analyzer

Benchmark App

Accuracy Checker

Aux. Capabilities

Traditional Computer Vision

Optimized Libraries & Code Samples

OpenCV*

OpenVX*

Samples

For Intel CPU & GPU/Intel® Processor Graphics

Tools & Libraries

Increase Media/Video/Graphics Performance

Intel® Media SDK
Open Source version

OpenCL™ Drivers & Runtimes

For GPU/Intel® Processor Graphics

Optimize Intel® FPGA (Linux* only)

FPGA RunTime Environment
(from Intel® FPGA SDK for OpenCL™)

Bitstreams

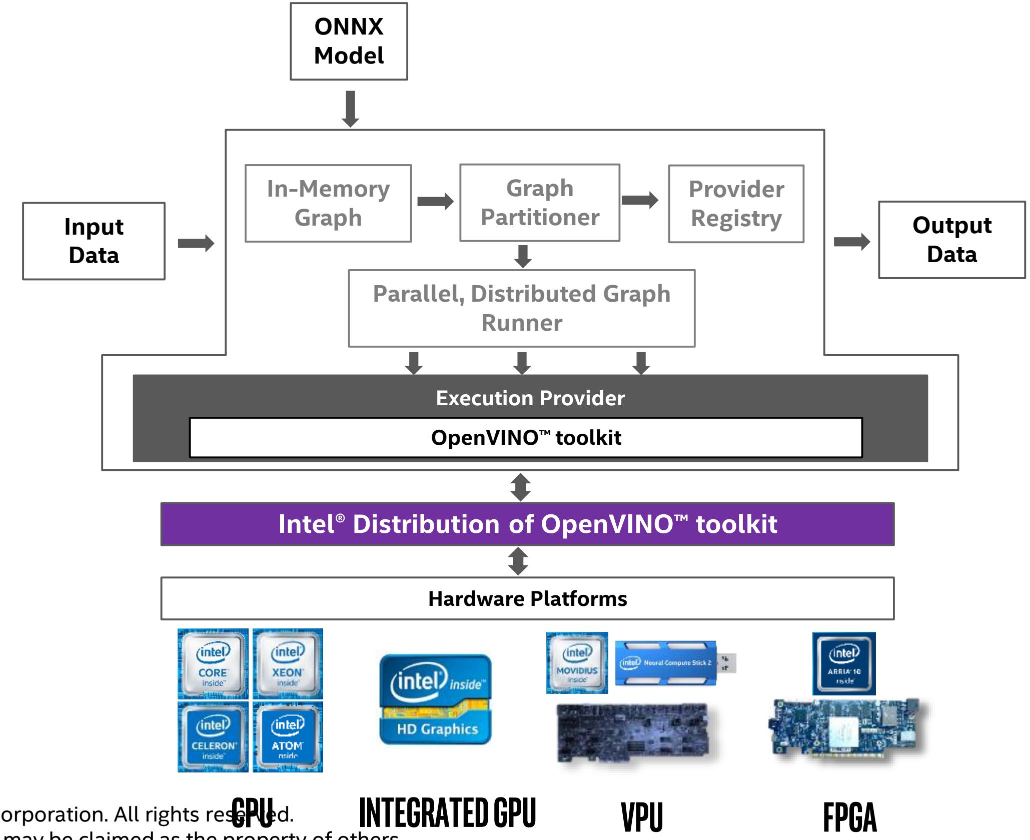
OS Support: CentOS* 7.4 (64 bit), Ubuntu* 16.04.3 LTS (64 bit), Microsoft Windows* 10 (64 bit), Yocto Project* version Poky Jethro v2.0.3 (64 bit), macOS* 10.13 & 10.14 (64 bit)

Intel® Architecture-Based Platforms Support

Intel® Vision Accelerator Design Products & AI in Production/ Developer Kits

An open source version is available at 01.org/openvinotoolkit (deep learning functions support for Intel CPU/GPU/NCS/GNA).

AT THE EDGE



ONNX Model

Input Data

In-Memory Graph

Graph Partitioner

Provider Registry

Parallel, Distributed Graph Runner

Output Data

Execution Provider

OpenVINO™ toolkit

Intel® Distribution of OpenVINO™ toolkit

Hardware Platforms

CPU

INTEGRATED GPU

VPU

FPGA

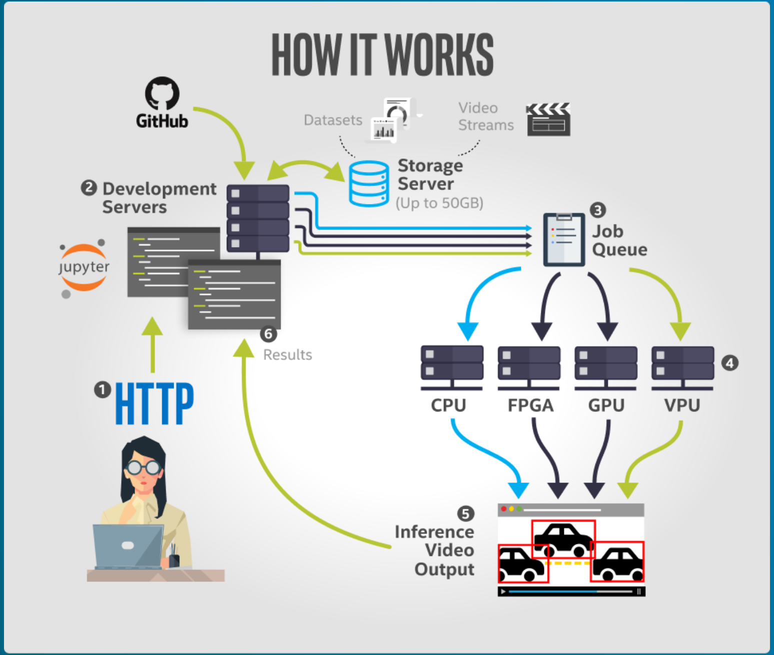
ONNX* ECOSYSTEMS

Frameworks	Caffe2 Chainer mxnet Microsoft Cognitive Toolkit ML.NET PaddlePaddle PyTorch MATLAB
Converters	ML LibSVM Keras TensorFlow scikit learn dmlc XGBoost
Runtimes	ONNX RUNTIME OpenVINO™ And others…
Visualizers	NETRON Visual DL

GET STARTED

Accelerate Time to Production with Intel® DevCloud for the Edge

See immediate AI Performance Across Intel's Array of Edge Solutions



- **Instant, Global Access**
  Run AI applications from anywhere in the world
- **Prototype on the Latest Hardware and Software**
  Develop knowing you're using the latest Intel technology
- **Benchmark your Customized AI Application**
  Immediate feedback - frames per second, performance
- **Reduce Development Time and Cost**
  Quickly find the right compute for your edge solution

Sign up now for access

3 SETUP OPTIONS IN GITHUB

BUILD FROM SOURCE

- (ONNX RT + OV) *
- RUN NATIVELY, COMPILE FROM SCRATCH – PROVIDES MAXIMUM FLEXIBILITY
- DEPLOY NATIVELY AT THE EDGE
- Github Readme ▸ https://github.com/microsoft/onnxruntime/blob/master/docs/execution_providers/OpenVINO-ExecutionProvider.md

DEPLOY CONTAINERS FROM AZURE IOT EDGE

- [ONNX RT + OV + AZURE IOT EDGE] *
- PROVIDES FLEXIBILITY AS WELL AS CONVENIENCE THROUGH CONTAINER SUPPORT
- DEPLOY CUSTOM APPLICATIONS IN CONTAINERS FROM AZURE IOT EDGE
- Github Azure IoT Hub Instructions ▸ https://github.com/intel/Edge-Analytics-FaaS/tree/master/Azure-IoT-Edge/OnnxRuntime

DEPLOY CONTAINERS FROM AZURE ML

- [ONNX RT + OV + AZURE IOT EDGE]
- MORE AUTOMATED, AZURE ML CONSTRUCTS THE CONTAINER FROM PRE-DEFINED AZURE ML FORMAT APPLICATIONS
- DEPLOY AZURE ML APPLICATIONS IN CONTAINERS FROM AZURE ML SERVICES
- Github Azure ML Container Dockerfiles ▸ https://github.com/microsoft/onnxruntime/tree/master/dockerfiles

*Note: Download the Intel® Distribution of OpenVINO™ toolkit installer(tgz) before building the above Docker image.

Additional Github Resource: Azure ML Instructions ▸

Cloud to Edge Deployment flow using Azure ML and Azure IoT Edge

Using Azure ML to deploy Azure ML container applications

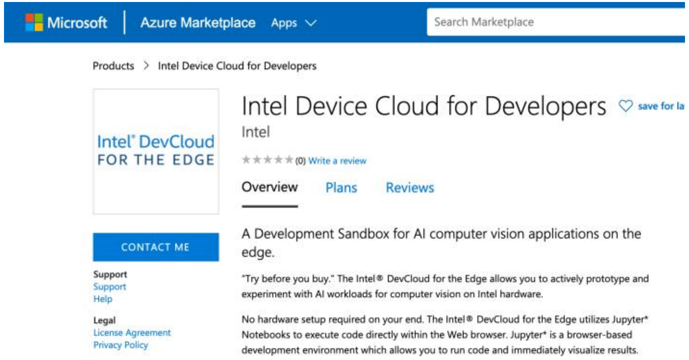
https://github.com/Azure-Samples/onnxruntime-iot-edge/tree/master/AzureML-OpenVINO

HOW IT WORKS (RUNTIME)

Simple runtime call pointing to model location

```
import onnxruntime


session = onnxruntime.InferenceSession(“model.onnx”)

x = GetInputData()

y = session.run([session.get_outputs()[0].name],

           {session.get_inputs()[0].name : x})
```

CSP MARKETPLACE OFFERS



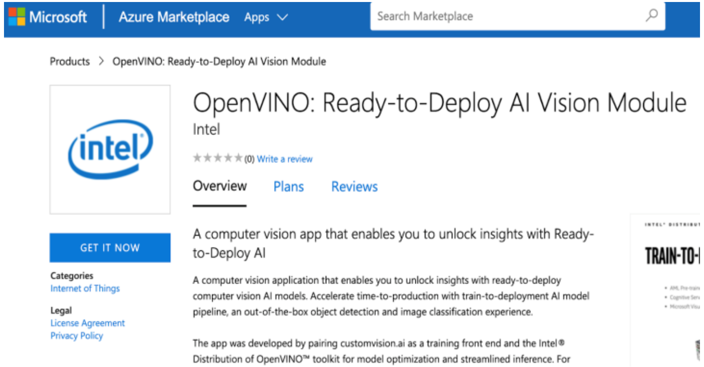
Intel Device Cloud for Developers

Intel

A Development Sandbox for AI computer vision applications on the edge.

"Try before you buy." The Intel® DevCloud for the Edge allows you to actively prototype and experiment with AI workloads for computer vision on Intel hardware.

No hardware setup required on your end. The Intel® DevCloud for the Edge utilizes Jupyter* Notebooks to execute code directly within the Web browser. Jupyter* is a browser-based development environment which allows you to run code and immediately visualize results.



OpenVINO: Ready-to-Deploy AI Vision Module

Intel

A computer vision app that enables you to unlock insights with Ready-to-Deploy AI

A computer vision application that enables you to unlock insights with ready-to-deploy computer vision AI models. Accelerate time-to-production with train-to-deployment AI model pipeline, an out-of-the-box object detection and image classification experience.

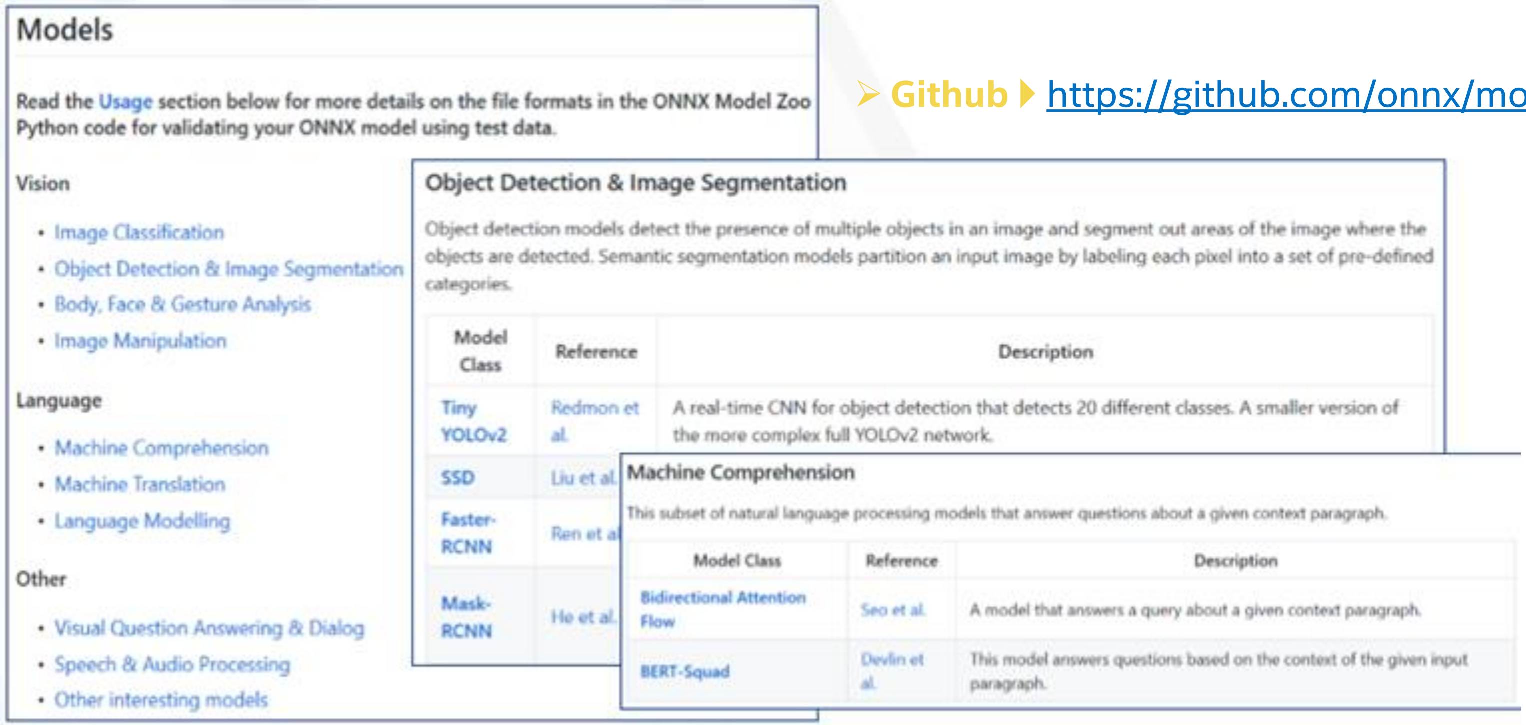
The app was developed by pairing customvision.ai as a training front end and the Intel® Distribution of OpenVINO™ toolkit for model optimization and streamlined inference. For

- Intel DevCloud link on Azure Marketplace
- Use Intel maintained HW Devices to quickly deploy models
- Understand HW device performance trade-off before you purchase your device(s).

- Vision Ready-to-deploy app link Azure Marketplace
- Leverages customvision.ai and OpenVINO with enhanced DX Training-to-Inference
- Bring your own Intel device(s) to experience the Cloud to Edge inference and quick relaunch of vision application

ONNX* MODEL ZOO

➢ Github ▸ https://github.com/onnx/models



Models

Read the Usage section below for more details on the file formats in the ONNX Model Zoo
Python code for validating your ONNX model using test data.

Vision

- Image Classification
- Object Detection & Image Segmentation
- Body, Face & Gesture Analysis
- Image Manipulation

Language

- Machine Comprehension
- Machine Translation
- Language Modelling

Other

- Visual Question Answering & Dialog
- Speech & Audio Processing
- Other interesting models

Object Detection & Image Segmentation

Object detection models detect the presence of multiple objects in an image and segment out areas of the image where the objects are detected. Semantic segmentation models partition an input image by labeling each pixel into a set of pre-defined categories.

Model Class	Reference	Description
Tiny YOLOv2	Redmon et al.	A real-time CNN for object detection that detects 20 different classes. A smaller version of the more complex full YOLOv2 network.
SSD	Liu et al.	
Faster-RCNN	Ren et al	
Mask-RCNN	He et al.	

Machine Comprehension

This subset of natural language processing models that answer questions about a given context paragraph.

Model Class	Reference	Description
Bidirectional Attention Flow	Seo et al.	A model that answers a query about a given context paragraph.
BERT-Squad	Devlin et al.	This model answers questions based on the context of the given input paragraph.

Sources: Microsoft

ONNX* TUTORIALS

➢ Github ▸ https://github.com/onnx/tutorials

Caffe2
mxnet
PyTorch
PaddlePaddle
Chainer
ML.NET
MathWorks
dmlc XGBoost
Cognitive Toolkit
ML
scikit learn
K

Sources: Microsoft

Get started with ONNX and tutorials

[Docker image for ONNX and Caffe2/PyTorch]

[Docker image for ONNX, ONNX Runtime, and various converters]

- **Getting ONNX models –** ONNX Model Zoo
- **Services –** Output ONNX models customized for your data
  - Azure Custom Vision service
  - Azure Machine Learning automated ML
- **Converting to ONNX format**
- **Scoring ONNX models –** Score accuracy

FEATURES SET

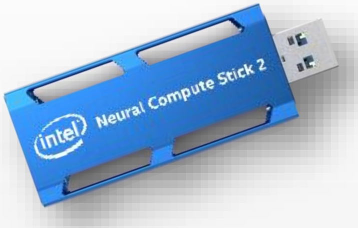
- Compute and Accelerator support:
  - Intel® CPU, integrated GPU
  - Intel® Movidius™ Myriad™ X VPU (USB and embedded)
  - Intel® Vision Accelerator Design Products with Intel® Movidius™ Myriad™ X VPU (2x, 4x & 8x)
  - Intel® Vision Accelerator Design Products with Intel® Arria® 10 FPGA
- Quantization support: Full precision (32 bit) and Half precision (16 bit) floating point
- Operator coverage: Majority models from ONNX Model Zoo github.com/onnx/models
- OS Support: Linux* and Win10*
- Docker container support: Linux* only
- Azure ML integration: Train model on Azure* ML and deploy on connected edge devices

DEVELOPER KITS, USE CASES, & CASE STUDIES

DEVELOPER KITS

INTEL® NEURAL COMPUTE STICK 2

link



Powered by the
Intel® Movidius™ Myriad™ X VPU

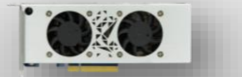
IEI TANK* AIOT DEVELOPER KIT

link

+

Intel® Vision Accelerator Design Product Choices

Powered by Intel® Movidius™ VPU (link)



Powered by Intel® Arria® 10 FPGA (link)

In Preview

UP SQUARED* AI VISION X DEV KIT

link

+

Intel® Vision Accelerator Design Product

Powered by Intel® Movidius™ VPU (link)

EQUIPMENT MAKER OFFERS

- IEI* TANK AIoT Developer Kit
  Intel® Core® i7/i5/i3 Processor & Intel® Xeon® Processor
  Use Case: Industrial
- IEI* FLEX-BX200
  Intel® Core® i3/5/i7 Processor
  Use Cases: Public Safety, Parking Mgmt., License Plate Detection

- UP* Squared; UP* Core Plus
  Intel® Atom™ Processor ; Intel® Core® i7/i5/i3 Processor
  Use Cases: Retail, DSS
- Aaeon* BOXER-6841M
  Intel® Core® i7/i5/i3 Processor
  Use Cases: Industrial, Smart Retail and Smart City

- Advantech* ARK-1124 + VEGA-320
  Intel® Atom™ Processor + Intel® Movidius™ Myriad™ X VPU
  Use Cases: Age & Gender Recognition

SUPPORT & RESOURCES

SUPPORT

Software Issues
Software issues related to ONNX Runtime with OpenVINO Execution Provider code should be logged at: “Issues” Tab https://github.com/Microsoft/onnxruntime with [OpenVINO-EP] tag.

Hardware Issues
Hardware issues should be routed towards your equipment maker suppliers, your Intel Representative, or Intel Premier Support

Supported Models
Link to supported models for the ONNX Runtime with OpenVINO Execution Provider https://github.com/microsoft/onnxruntime/blob/master/docs/execution_providers/OpenVINO-ExecutionProvider.md.
All issues related to these models should be routed towards your Intel Representative

Intel® Distribution of OpenVINO™ toolkit Support
OpenVINO issues should be reported through the OpenVINO “Computer Vision” Forum https://software.intel.com/en-us/forums/computer-vision

ONNX Support
All other ONNX model issues should be logged at “Issues” tab https://github.com/Microsoft/onnxruntime

INTEL® IOT RFP READY KITS

Check RFP Ready Kit Playbook for details on each kit

Retail

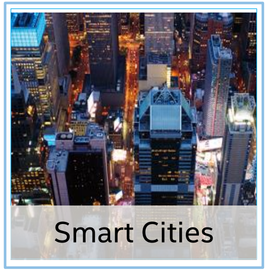
- Kiosk & digital signage
- POS & mobile POS
- Inventory Management

Industrial

- Manufacturing
- Building Management
- Agriculture
- Energy

Transportation

- Fleet Management
- Logistics



Smart Cities

- Security surveillance
- Smart lighting
- Connected Transportation
- Air quality management

Healthcare

- Medical (in-hospital)
- Remote health management

Horizontal

- Security Surveillance Video
- Connectivity

ACCELERATE PROTOTYPE TO PRODUCTION & SOLUTION DEPLOYMENT

DEVELOP ON HOST SYSTEM

& TEST on

INTEL DEVCLOUD FOR THE EDGE

OpenVINO™

USE VISION ACCELERATOR KITS

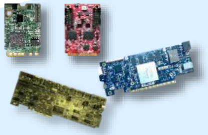
Intel® Distribution of OpenVINO toolkit

AAEON UP Squared AI Vision Developer Kit

IEI Tank AIoT Developer Kit

INCREASE PERFORMANCE

Intel® Vision Accelerator Design Products



Intel® Movidius™ Myriad™ X VPU
Intel® Arria® 10 FPGA

DEVELOP USE CASE SPECIFIC OFFERS

Developer optimization & use case specific applications

Intel® RFP Ready Kits

SCALE

Deploy solution & solve business problems, and scale with Intel® IoT Solution Aggregators & Ecosystem

intel market ready ✓

Intel® AI: In Production ▶

https://software.intel.com/ai-in-production intel®