エッジAI向けOSS動向および NXPのAI開発への取り組みのご紹介

NXPジャパン株式会社
アプリケーション技術統括部
上釜 悠聖

JULY 2022

SECURE CONNECTIONS FOR A SMARTER WORLD

Agenda

1. 会社紹介
2. エッジAI向けOSS動向
3. エッジAI開発におけるポイント
4. NXPのAI開発への取り組み(eIQ)
5. まとめ

1. 会社紹介

SECURE CONNECTIONS
FOR A SMARTER WORLD

NXPセミコンダクターについて

30+カ国
本社：オランダ
(アイントホーフェン)

60+
60年以上の歴史

約31,000
従業員数

約11,000
エンジニア数

9,500
特許数

$11.06 B
年間売上高（注1）

注1 ：2021会計年度の実績：その他の情報については、 NXPウェブサイトのInvestor Relations セクション（www.nxp.com/investor）にあるFinancial Informationページをご覧ください。

フル・システム・ソリューションを提供

ワールドクラスのコネクティビティ・ポートフォリオ

Wi Fi
UWB
Bluetooth
ZigBee
MIFARE
MiGLO
eSIM
5G
DSRC
（V2X）
RAIN RFID
sigfox
THREAD GROUP

ユニークなMCU / MPU製品ポートフォリオ

i.MX 6 / 7 / 8 / 8M
高性能、3Dグラフィックス

S32車載プラットフォーム
スケーラブル、セキュア、OTA、ASIL D

Layerscape™
高速イーサネット、TSN

i.MX RT
最大1GHzの高性能MCU

Kinetis™ / LPC
低コストから高集積製品まで

信頼性の高いソリューションを付加

- AI/ML ソリューション
- IoT向けセキュア・エレメント
- ソフトウェアエコシステム
- セキュアな車載ソリューション

幅広いお客様にフルシステムソリューションを提供

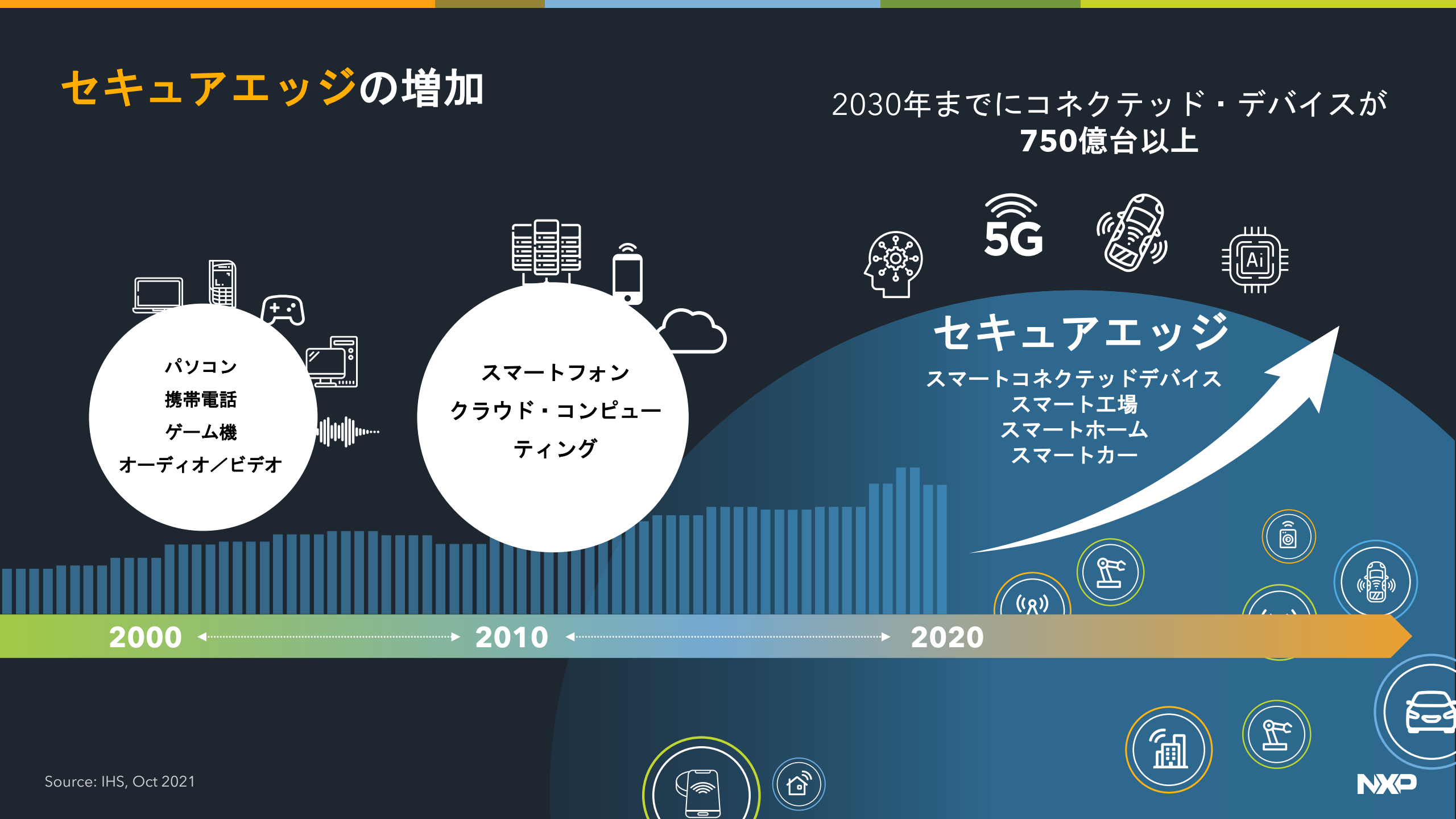
- オートモーティブ
- インダストリアル
- モバイル
- 通信インフラ

2. エッジAI向けOSS動向

SECURE CONNECTIONS
FOR A SMARTER WORLD

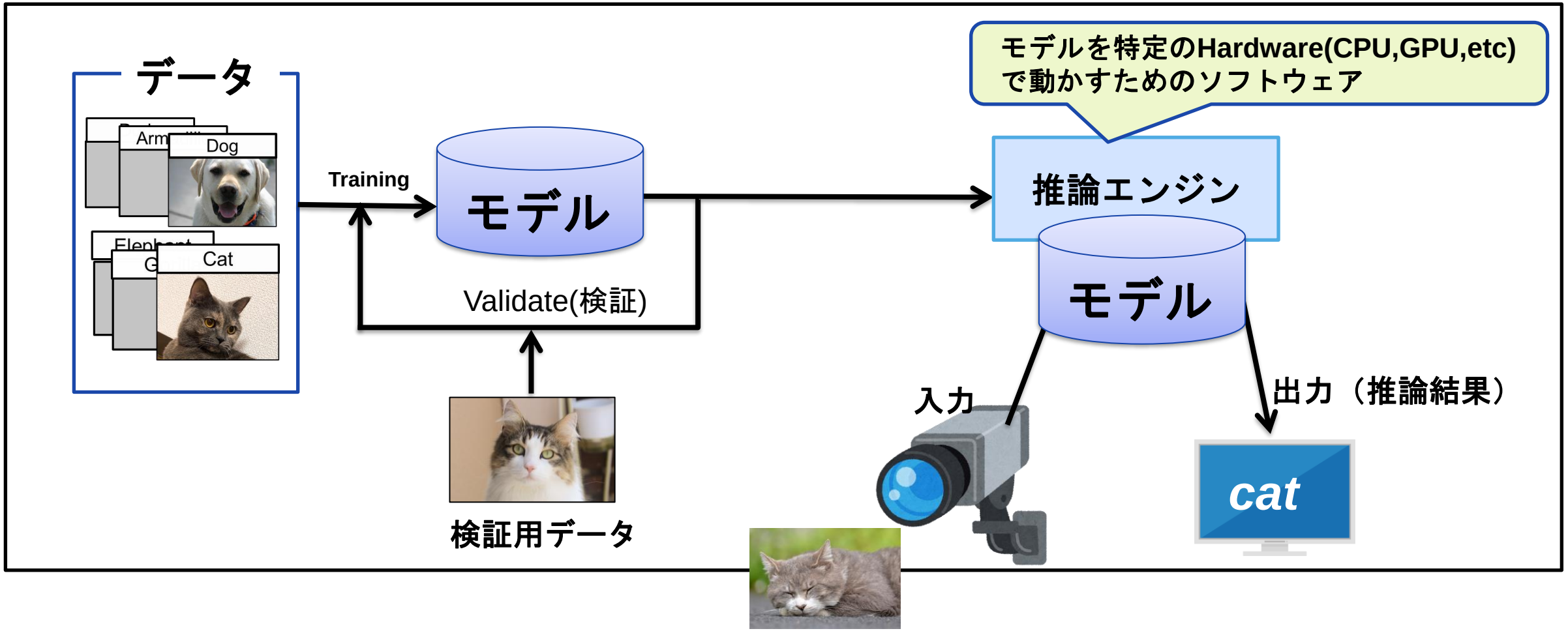
セキュアエッジの増加

2030年までにコネクテッド・デバイスが
750億台以上



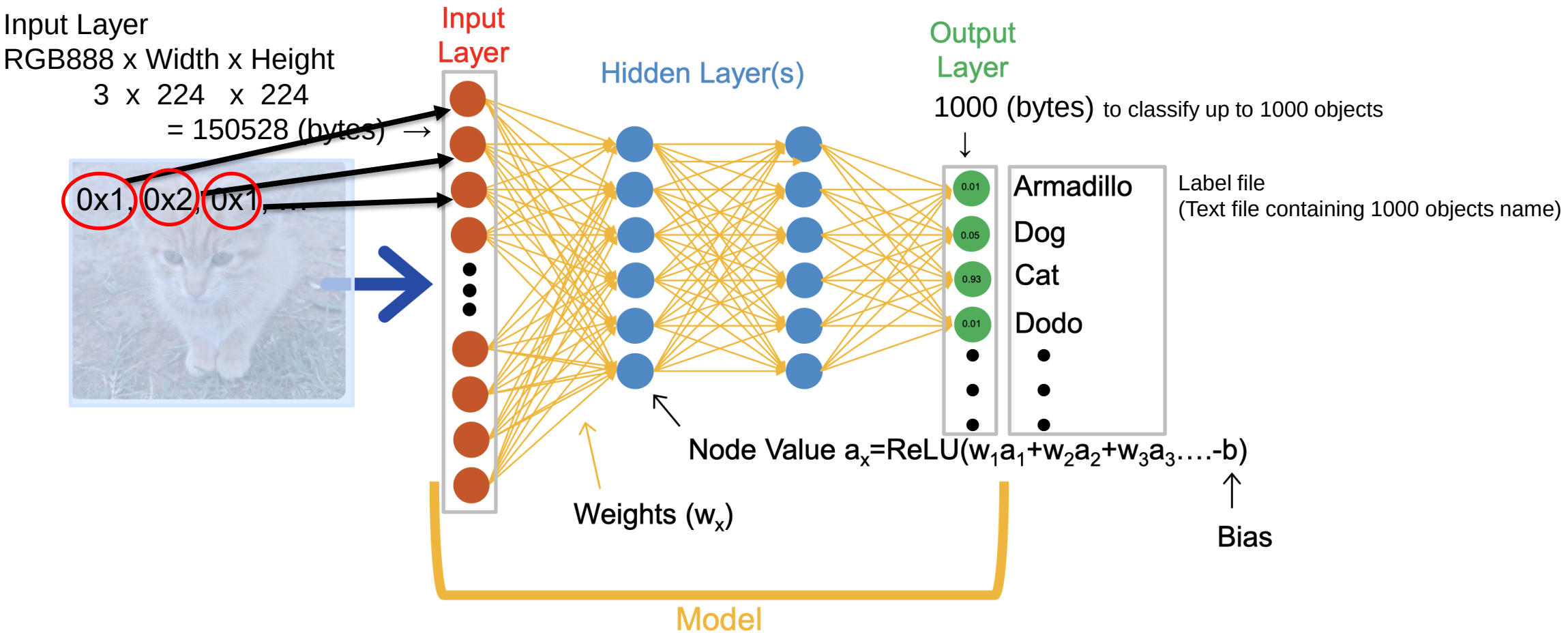
Source: IHS, Oct 2021

AI(Deep Learning)開発フロー 概要



Deep Learningモデル構造の概要

- Example： Mobilenet (NN model for Image Classification)



mobilenet_v1_1.0_224_quant.tflite

Deep Learning Frameworkを
用いた開発が主流

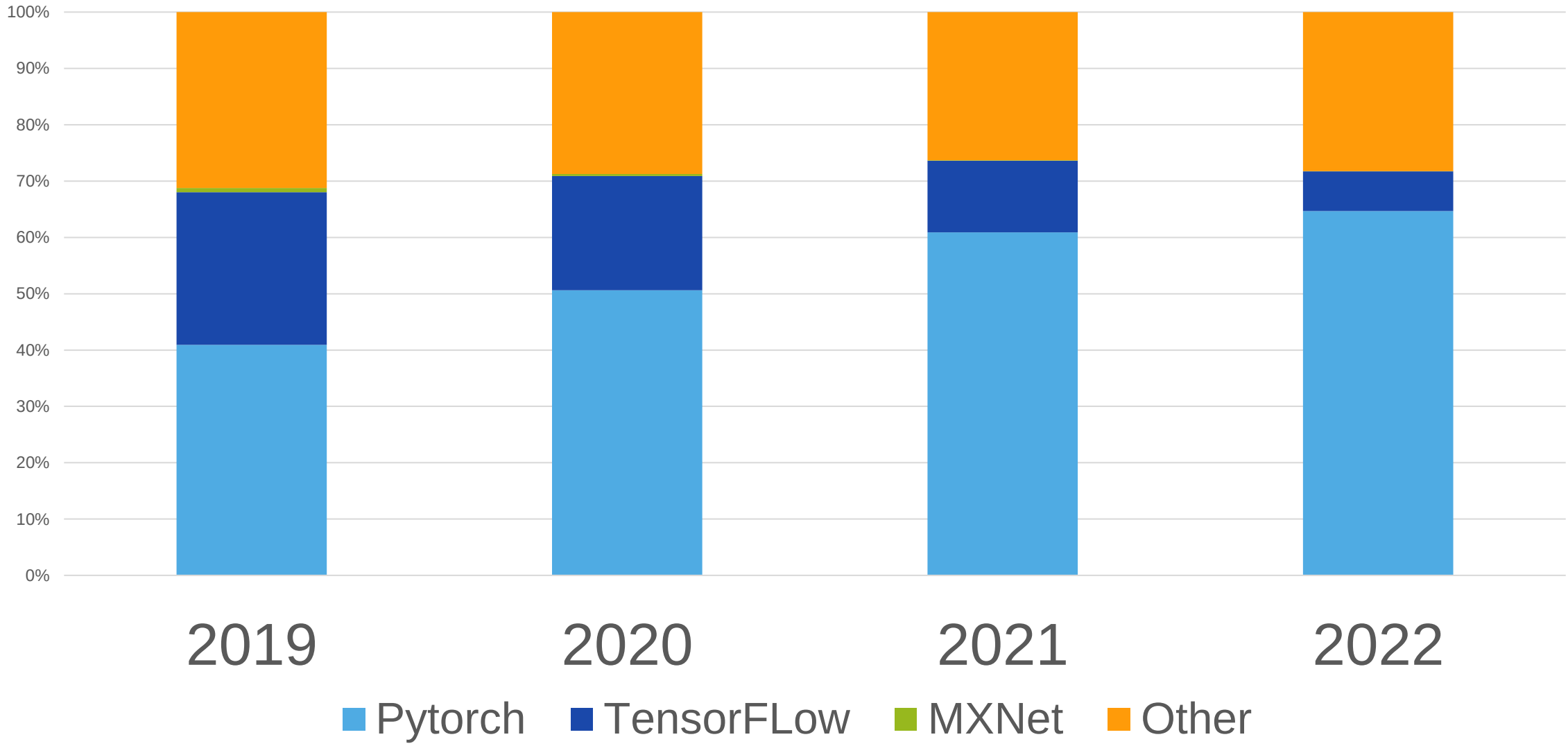
様々なDeep Learning Frameworkが存在
どれから勉強すればいいのか・・・

Deep Learning Frameworkの動向を調査！

人気があるOSSのDeep Learning Framework

2020	2021	2022
TensorFlow	**TensorFlow**	**TensorFlow**
Pytorch	**Pytorch**	**Pytorch**
Apache MXNet	**Apache MXNet**	**Apache MXNet**
Apache PredictionIO	Caffe2	Keras
Caffe2	Chainer	Deeplearning4j
Chainer	CTNK	
CTNK	Deeplearning4j	
Neon	Theano	
PaddlePaddle	Torch7	

論文でのDeep Learning Framework使用比率



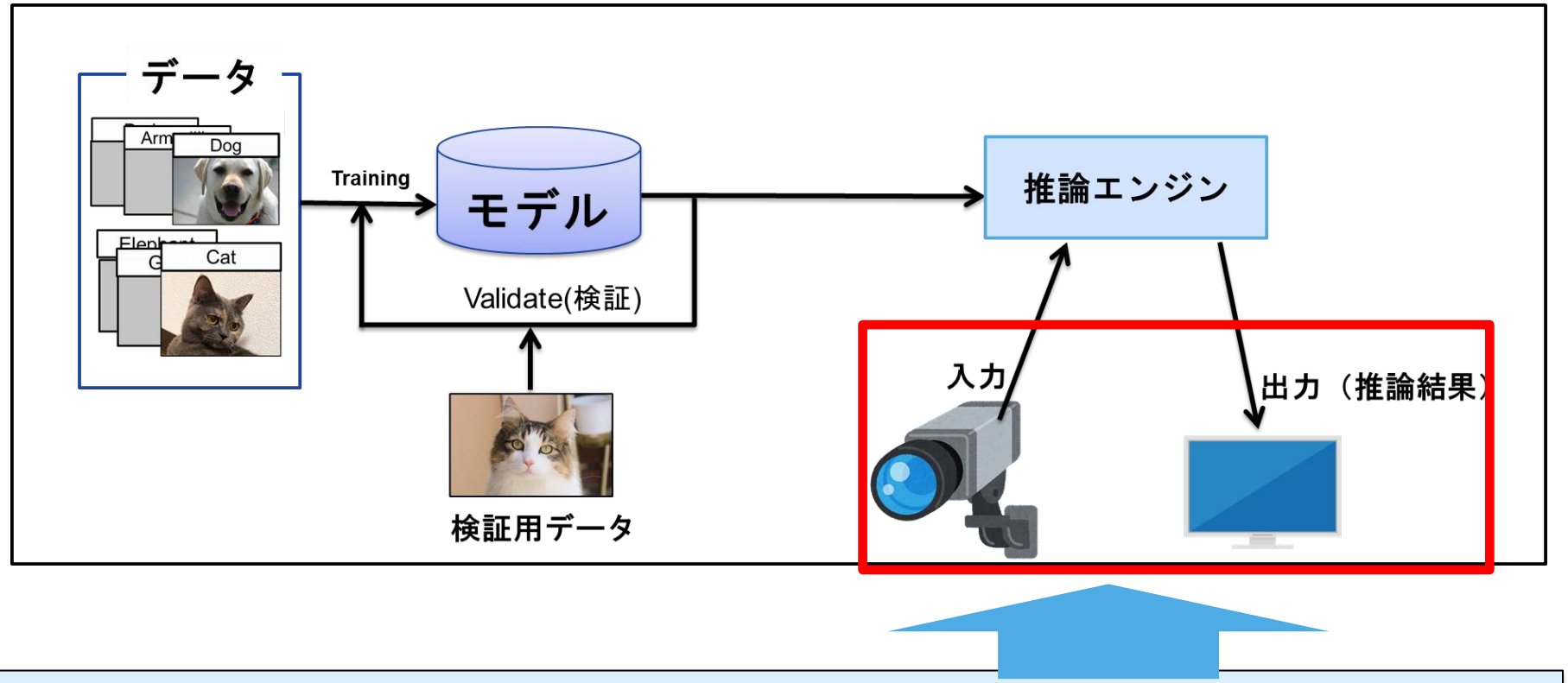
Define-By-RunとDefine-And-Run

	Define-By-Run	Define-And-Run
方式	ネットワークの 構築と実行が**同時**	ネットワークを 構築**後**に実行
メリット	デバッグしやすい	最適化しやすい
Framework	Pytorch, Chainerなど	TensorFlow, MXNet など

主要なOSSのDLフレームワークの特徴

名称	開発元	対応言語	特徴
TensorFlow	Google	Python, C++ Java, Go	利用者が多い、Githubでも人気が高い Keras(高レベルAPI）で実装しやすい
Pytorch	Facebook	Python	研究分野での人気が高い Define-by-Run、デバッグしやすい
Apache MXNet	ワシントン大学, カーネギーメロン大学	Python,C++ Scala,R Matlab, Julia, etc	Define-and-Run 対応言語多 AWSが公式サポートを表明

DLアプリ(vision)開発で使われるOSS（マルチメディア系）



マルチメディア OSS

Gstreamer

NNStreamer

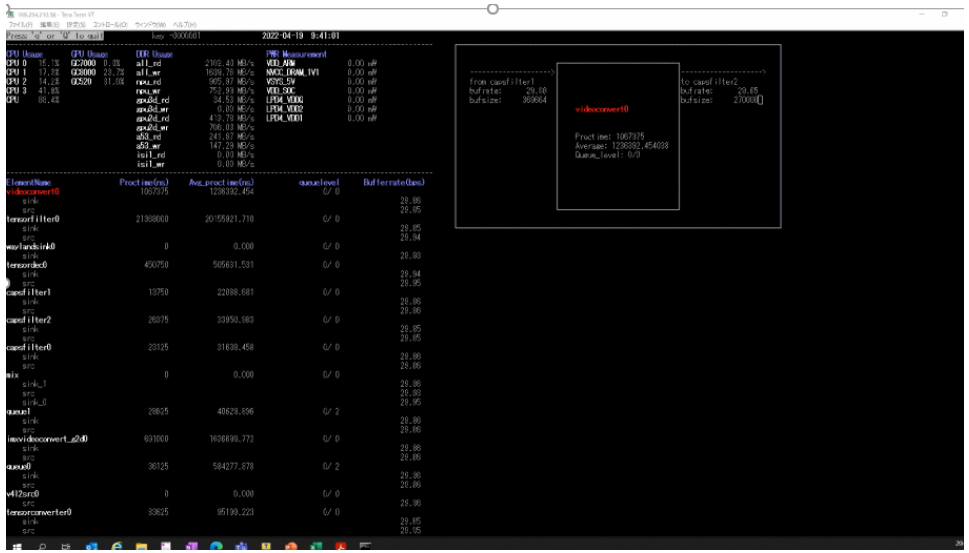
OpenCV

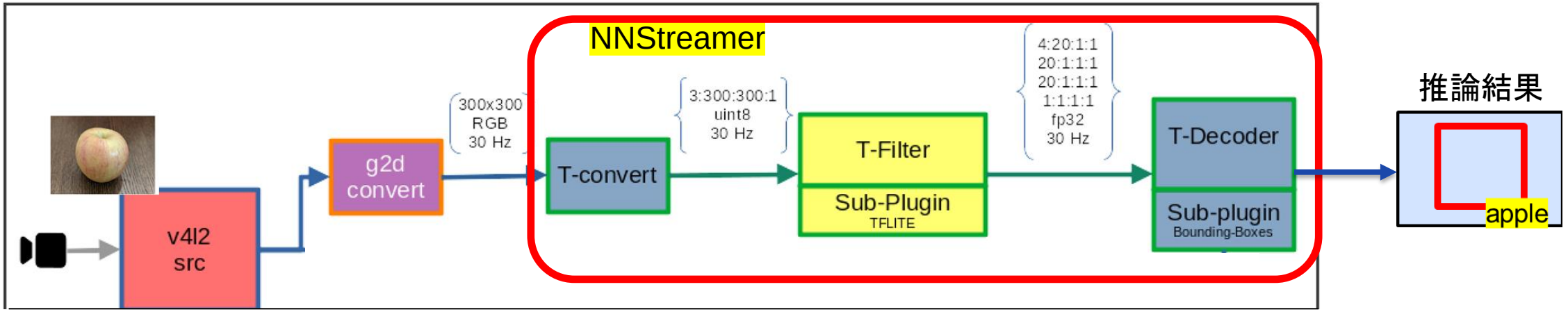
CV:DNN

etc

NNStreamerとは

- GstreamerのML処理プラグイン
- Samsungが開発⇒LF AI Foundationに移管
- NNSharkでリアルタイムプロファイル





3. エッジAI開発におけるポイント

SECURE CONNECTIONS
FOR A SMARTER WORLD

エッジデバイスでAI処理を行うメリットと課題

メリット

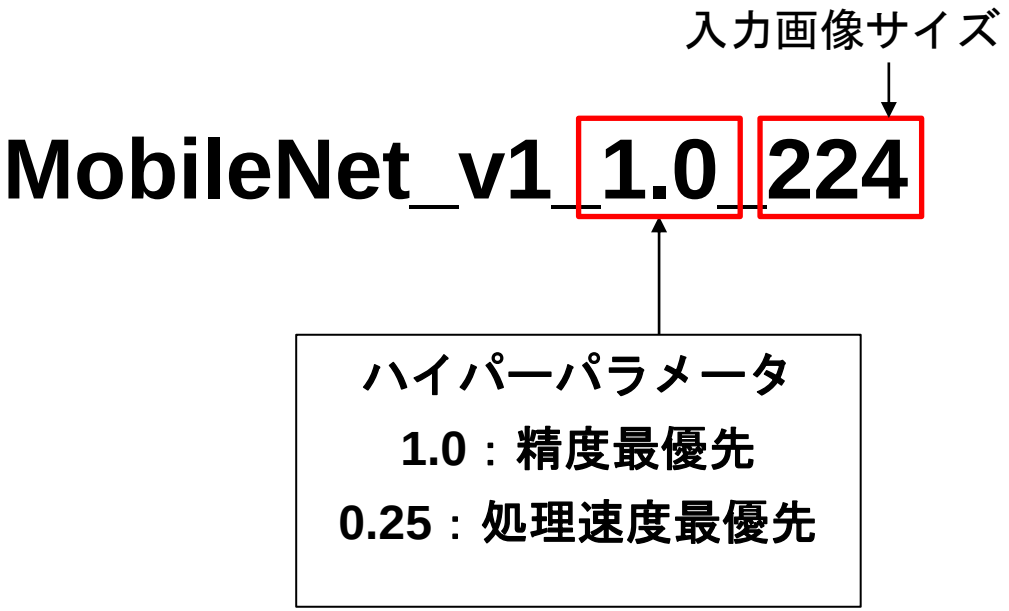
1.セキュリティの向上
2.リアルタイム性の向上
3.コスト削減

課題

1.Hardwareリソースに限りがある
2.学習と異なる環境で推論を実施する場合が多い

モデルの精度と処理負荷はトレードオフの傾向

Model	Million MACs	Million Parameters	Top-1 Accuracy
MobileNet_v1_1.0_224	569	4.24	70.9
MobileNet_v1_1.0_192	418	4.24	70.0
MobileNet_v1_1.0_160	291	4.24	68.0
MobileNet_v1_1.0_128	186	4.24	65.2
MobileNet_v1_0.75_224	317	2.59	68.4
MobileNet_v1_0.75_192	233	2.59	67.2
MobileNet_v1_0.75_160	162	2.59	65.3
MobileNet_v1_0.75_128	104	2.59	62.1
MobileNet_v1_0.50_224	150	1.34	63.3
MobileNet_v1_0.50_192	110	1.34	61.7
MobileNet_v1_0.50_160	77	1.34	59.1
MobileNet_v1_0.50_128	49	1.34	56.3
MobileNet_v1_0.25_224	41	0.47	49.8
MobileNet_v1_0.25_192	34	0.47	47.7
MobileNet_v1_0.25_160	21	0.47	45.5
MobileNet_v1_0.25_128	14	0.47	41.5



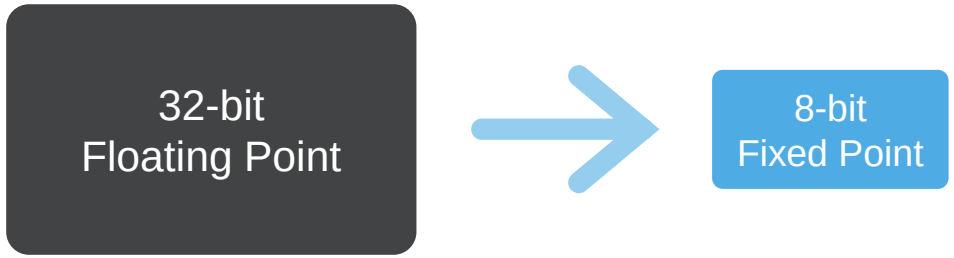
「計算量・ネットワークサイズ」と
「Accuracy(精度)」はトレードオフの関係性

表：代表的な画像分類モデルMobileNet_v1

モデルの最適化

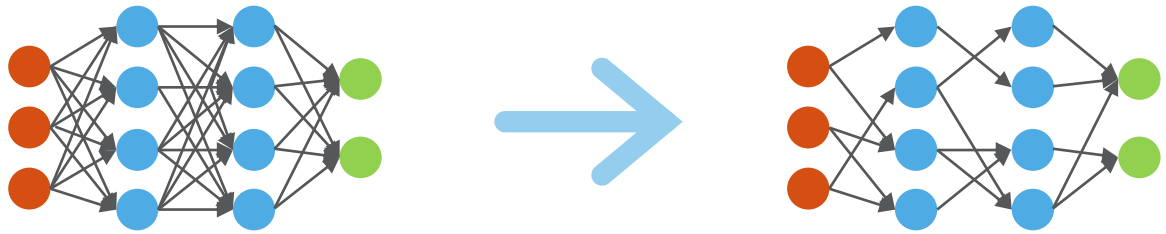
Quantization(量子化)

- データ型を32bit浮動小数点から8bit整数に変換
  - モデルサイズを1/4に削減
  - 整数処理は浮動小数点処理よりも高速
  - 適切な処理で行えばaccuracy(精度)のロスは少ない

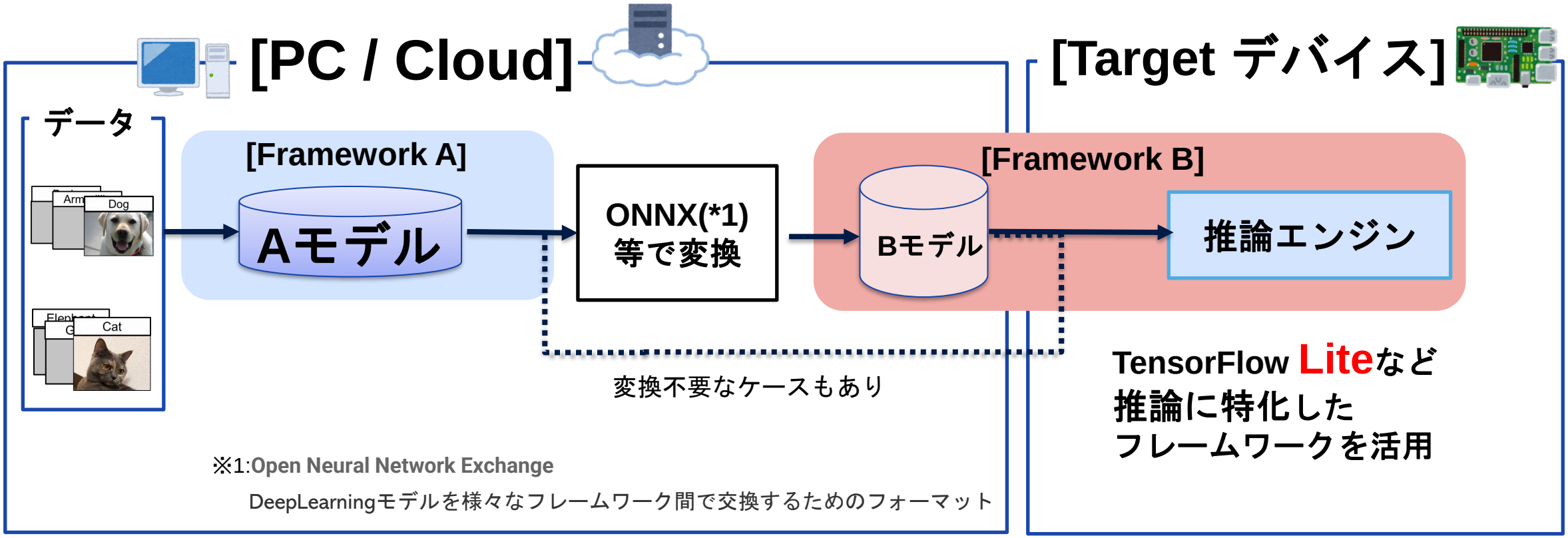


Pruning （枝刈り)

- 重要でないネットワークノードを削除する
  - Pruning実施後は再度トレーニングを行うことが推奨されます



学習異なる環境で推論⇒モデル変換



PC/クラウドにて好きなフレームワークでモデルを作成

⇒TargetデバイスのBSPが対応するFrameworkに変換してデプロイ

　推論に特化したフレームワークを使用する形式が一般的

4. NXPのAI開発への取り組み(eIQ)

SECURE CONNECTIONS
FOR A SMARTER WORLD

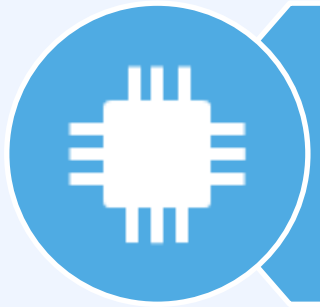
eIQ機械学習ソフトウェア開発環境とは

NXP製MCU/MPU向けMachine Learning(機械学習)アプリを構築するための
ソフトウェアライブラリおよび開発ツール一式

eIQ特徴

Ready To start

・テスト及び最適化実施済みの推論エンジン
・すぐにエッジAIを試せるサンプルソフトウェア



スケーラビリティ

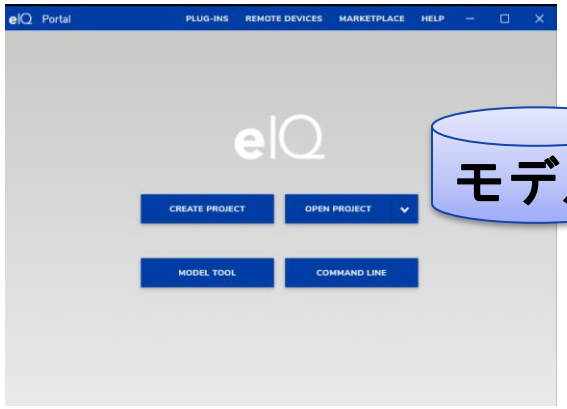
i.MX 8M シリーズ全て、
i.MX RTシリーズの一部に対応

・モデルトレーニングツール
・無償利用可能

eIQ 機械学習ソフトウェア開発環境

eIQ

[Host PCで動作]
eIQ Portal



モデル

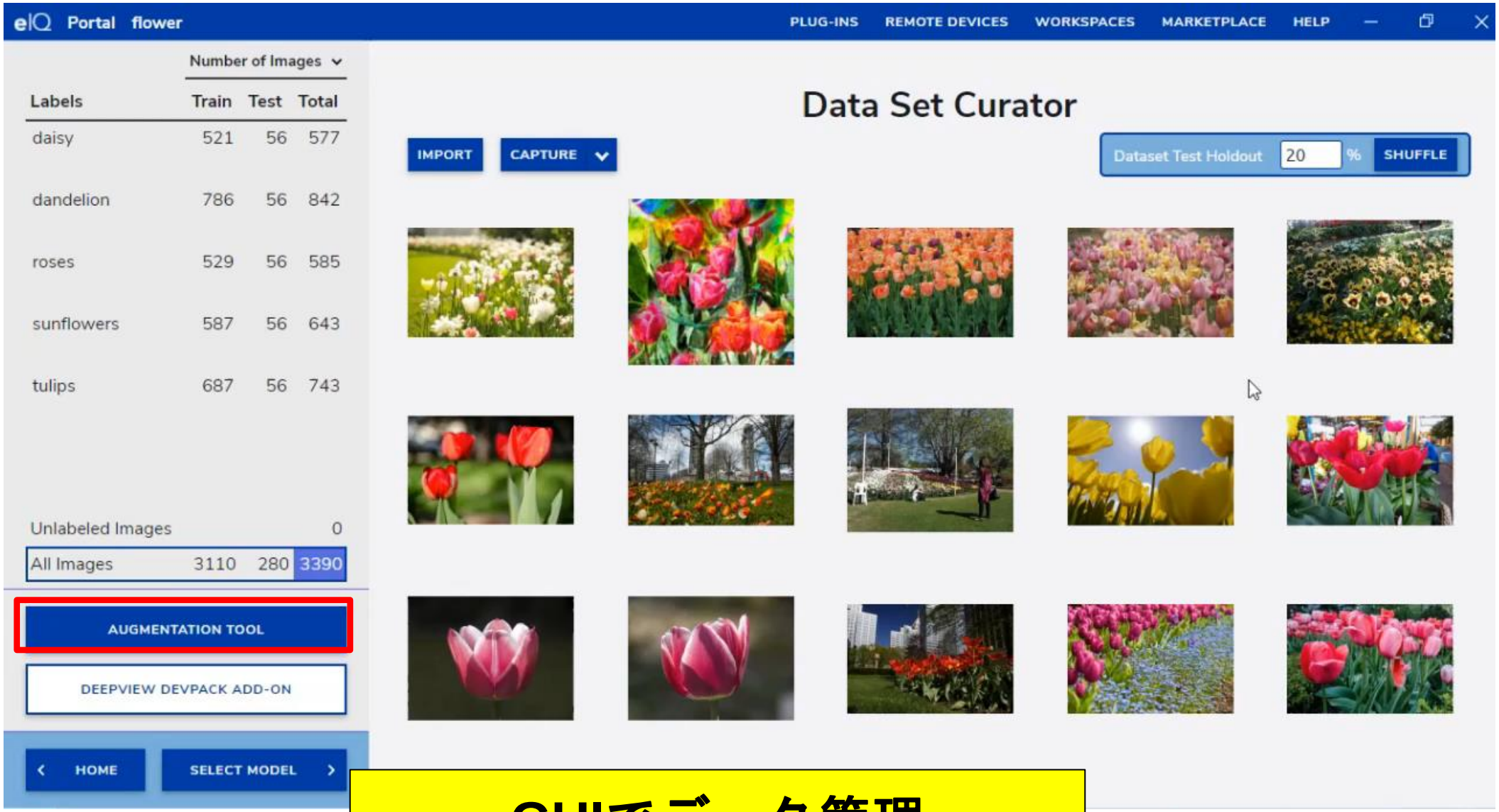
デプロイ

- モデルトレーニング
- モデル評価
- モデル変換,最適化　etc

[MCU/MPUで動作](BSPに統合)

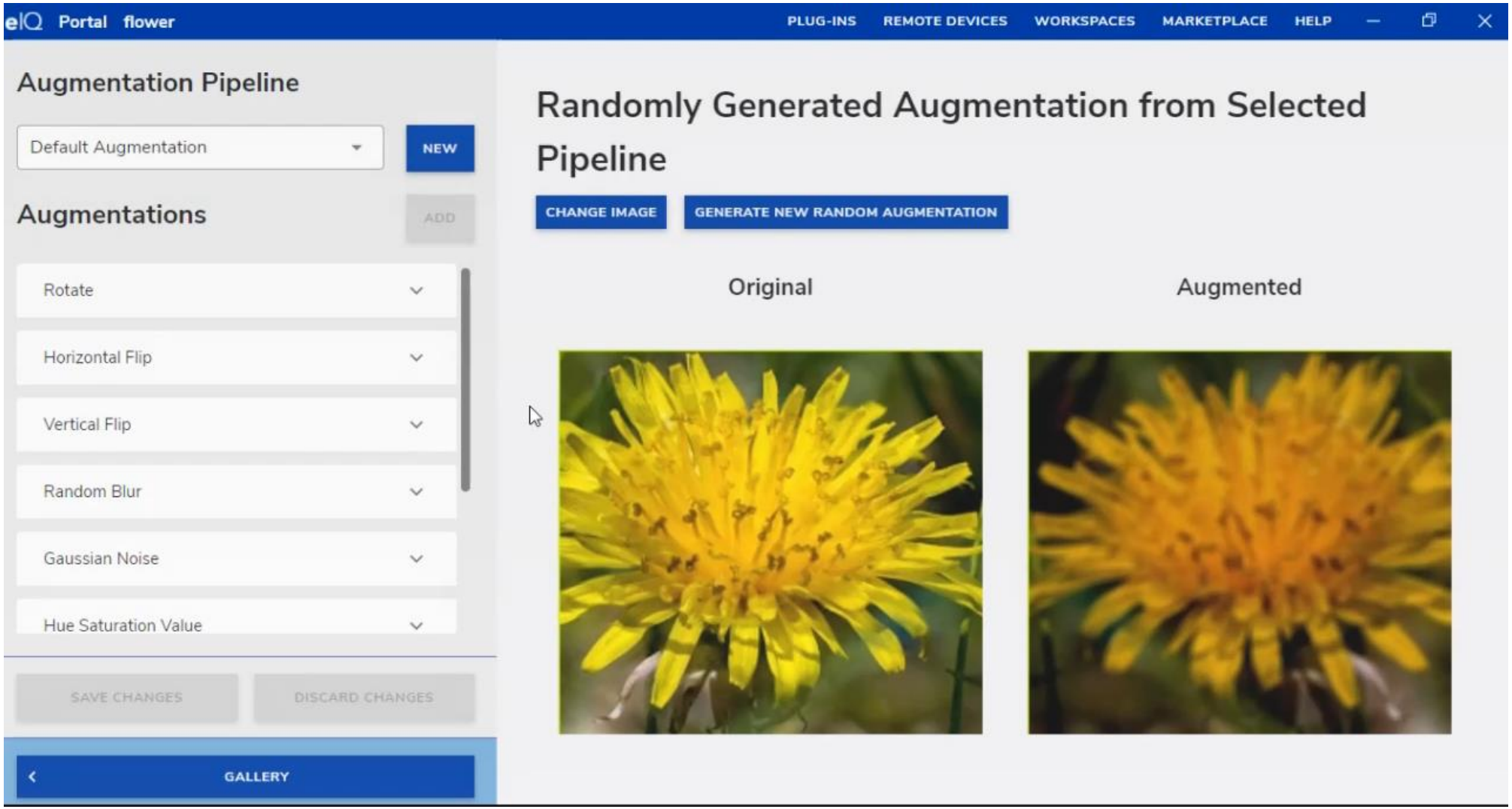
Sample Application	Object Detection	Object Classification	etc	
Inference Engines	Tensor Flow Lite	Pytorch	DeepView RT	OpenCV
	Tensor Flow Micro	ONNX Runtime	Arm® NN	Glow
Hardware Abstraction Layer	Open VX	NNRT	ACL	etc
↑ S/W				
↓ H/W				
Compute Unit	Arm® Cortex®-A	Arm® Cortex®-M	GPU	DSP	NPU

eIQ Portal紹介(データ管理)



GUIでデータ管理

eIQ Portal紹介(Training用データ拡張)

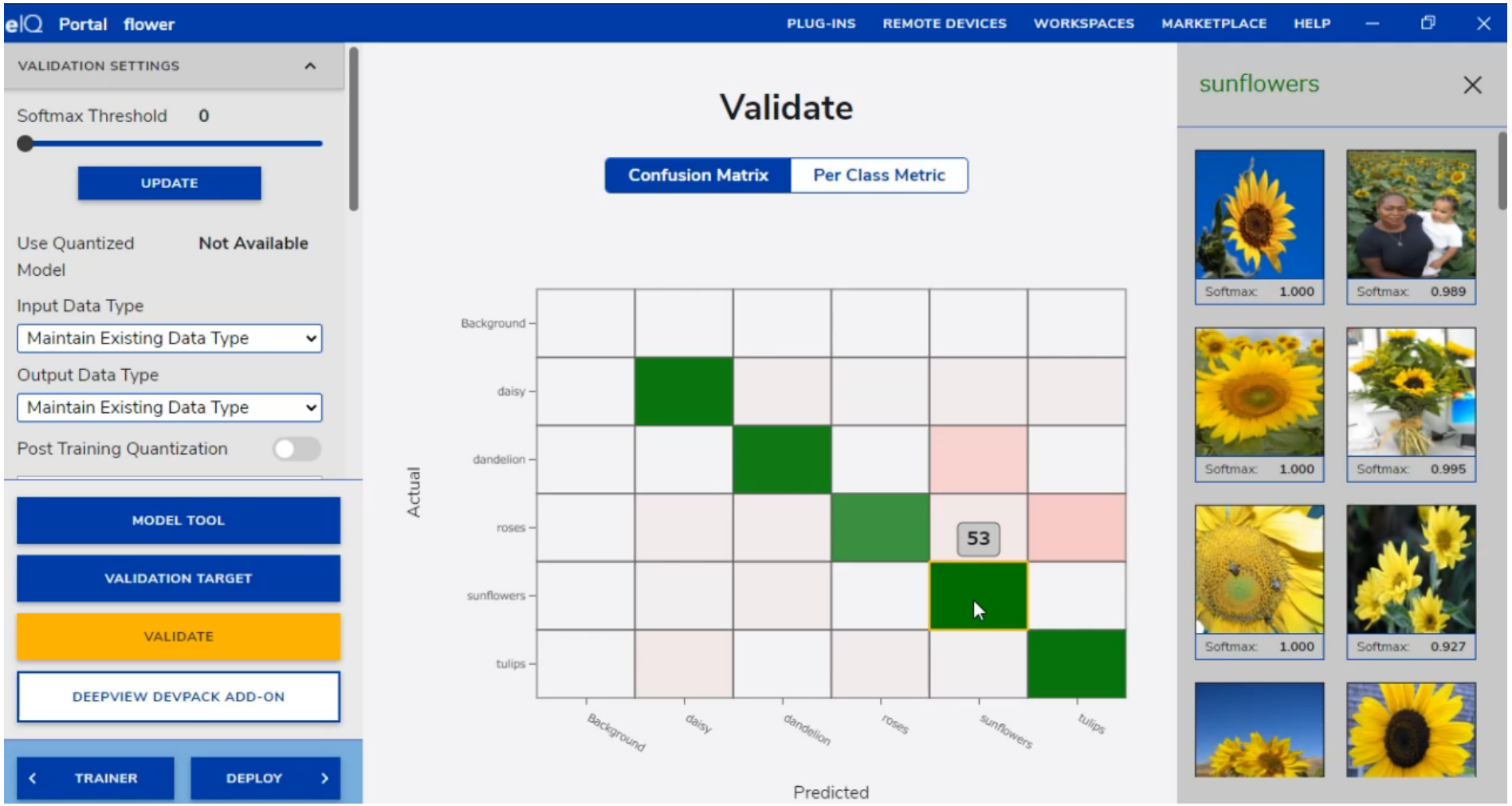


データを加工⇒学習用データを水増し

eIQ Portal紹介(トレーニング設定)

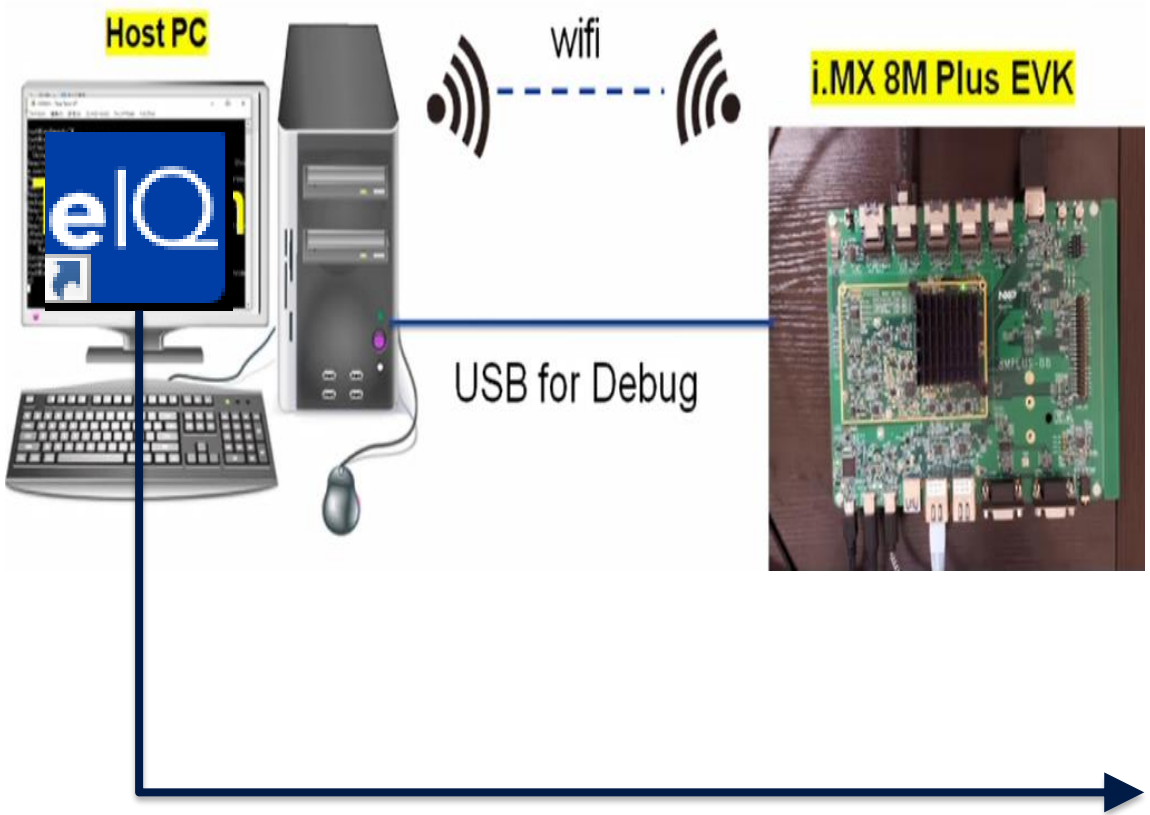
eIQ Portal紹介(トレーニング)

eIQ Portal紹介(検証)

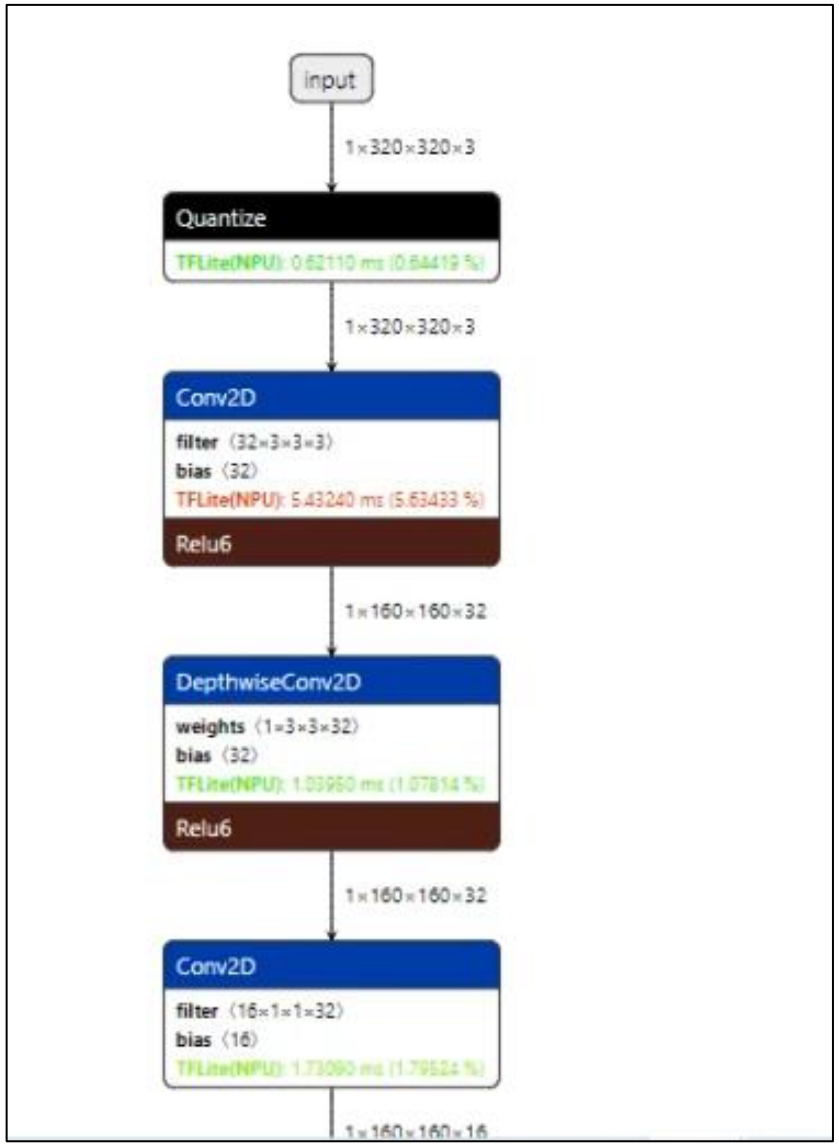


検証結果をGUIで確認可

ターゲットデバイス上でプロファイル(GUI)

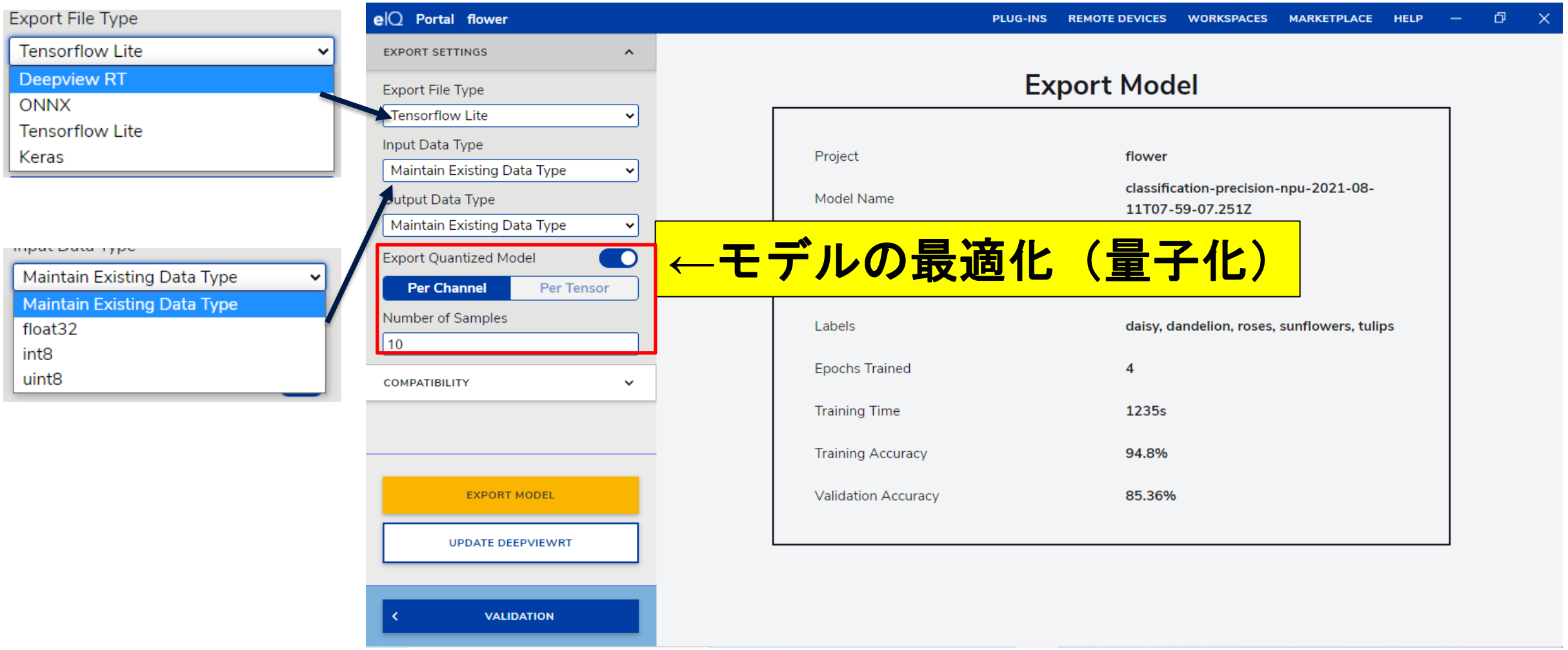


作成したモデルをターゲットデバイスでプロファイル

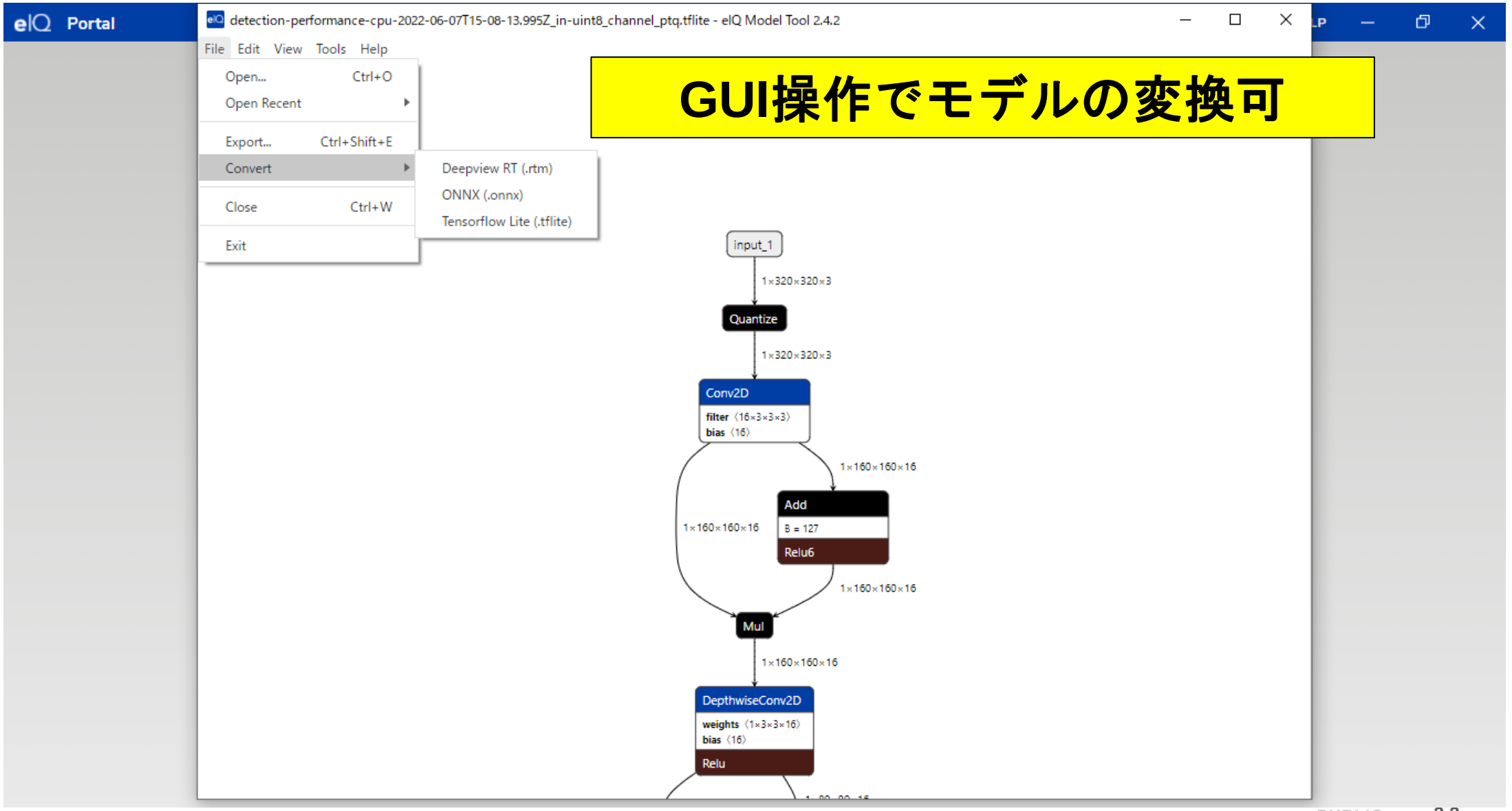


プロファイル結果をGUIで確認可

eIQ Portal紹介(Export)



モデル変換機能


eIQ Portal
detection-performance-cpu-2022-06-07T15-08-13.995Z_in-uint8_channel_ptq.tflite - eIQ Model Tool 2.4.2
File Edit View Tools Help
Open... Ctrl+O
Open Recent
Export... Ctrl+Shift+E
Convert
Close Ctrl+W
Exit
Deepview RT (.rtm)
ONNX (.onnx)
Tensorflow Lite (.tflite)
GUI操作でモデルの変換可
input_1
1×320×320×3
Quantize
1×320×320×3
Conv2D
filter 〈16×3×3×3〉
bias 〈16〉
1×160×160×16
Add
B = 127
Relu6
1×160×160×16
1×160×160×16
Mul
1×160×160×16
DepthwiseConv2D
weights 〈1×3×3×16〉
bias 〈16〉
Relu

i.MXプロセッサ＋eIQによるデモ

SECURE CONNECTIONS
FOR A SMARTER WORLD

eIQ対応MPU/MCUラインナップ

MPU

i.MX 8Mアプリケーションプロセッサシリーズ(Cortex®-A53)

i.MX 8M
- *4x/2x Cortex®-A53 @ 1.5GHz*
- *Cortex®-M4 @ 266MHz*
- *GPU*

i.MX 8M Nano
- 4x/2x/1x Cortex®-A53 @ 1.6GHz
- Cortex®-M7 @ 750MHz
- GPU

i.MX 8M Mini
- 4x/2x/1x Cortex®-A53 @ 1.8GHz
- Cortex®-M4 @ 400MHz
- GPU

i.MX 8M Plus
- 4x/2x Cortex®-A53 @ 1.8GHz
- Cortex®-M7 @ 800MHz
- GPU
- **NPU (2.3TOPS)**

performance

Software Compatible

MCU

i.MX RTクロスオーバー・MCUシリーズ(Cortex-M33, M7)

i.MX RT600
- *Cortex®-M33@300MHz*
- *DSP@600MHz*

i.MX RT1050/1060
- Cortex®-M7@600MHz

i.MX RT1170
- Cortex®-M7@1GHz
- Advanced Security

performance

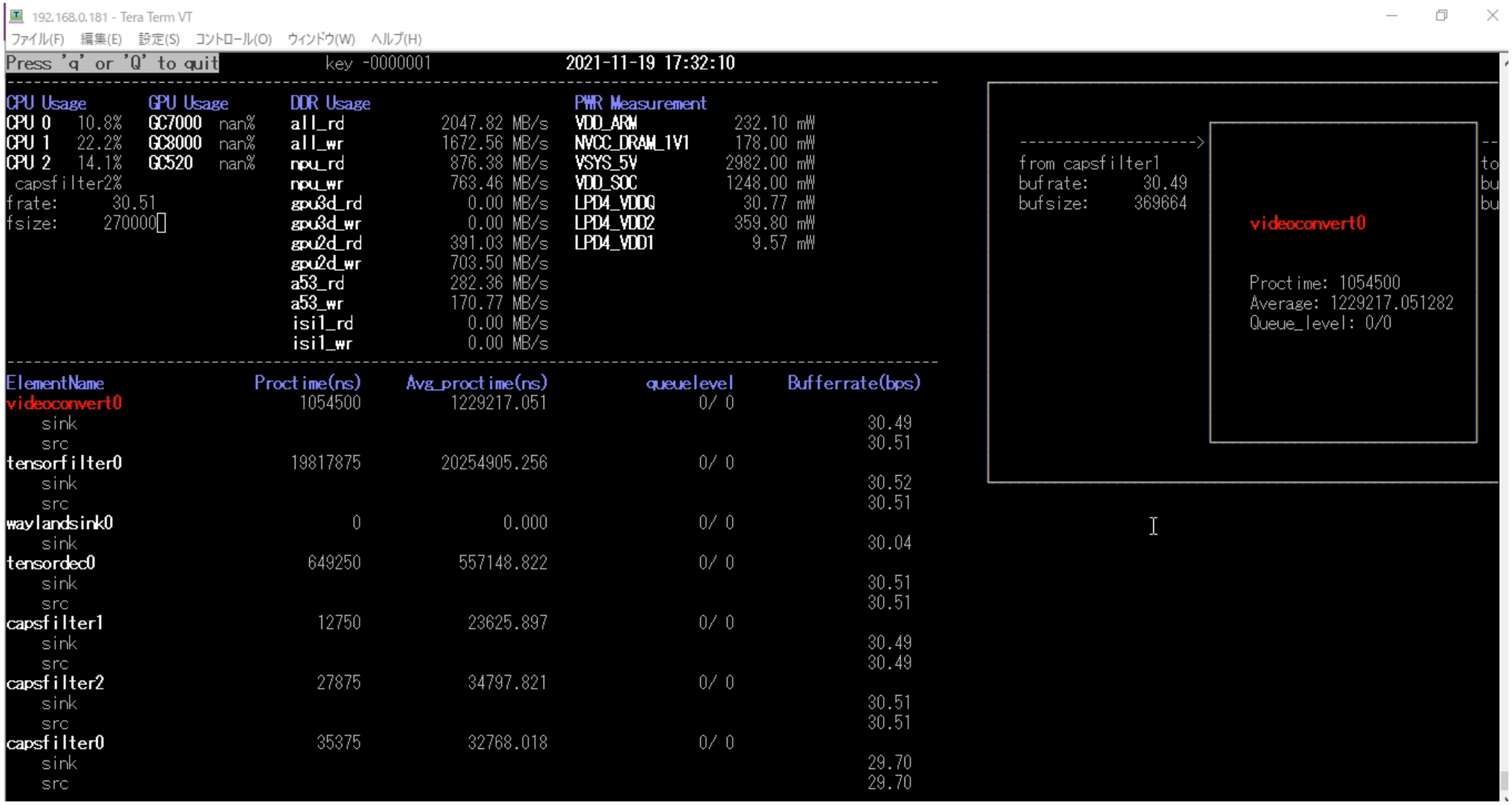
デモ①NNSTREAMERを用いたOBJECT DETECTION

COM10 - Tera Term VT

ファイル(F) 編集(E) 設定(S) コントロール(O) ウィンドウ(W) ヘルプ(H)

```
root@imx8mpevk:~/nnstreamer#
root@imx8mpevk:~/nnstreamer# gst-launch-1.0 --no-position v4l2src device=/dev/vi
deo2 ! video/x-raw, width=640,height=480,framerate=30/1 ! tee name=t t. ! queue
e-buffers=2 ! leaky=2 ! imxvideoconvert_g2d ! video/x-raw,width=300,height=300,f
ormat=RGBA ! videoconvert ! video/x-raw, format=RGB ! tensor_converter ! tensor_
framework=tensorflow-lite model=${MODEL} custom=Delegate:NNAPI ! tensor_decoder
mode=bounding_boxes option1=tf-ssd option2=${LABELS} option3=0:1:2:3,50 option4=
 option5=300:300 ! mix. t. ! queue max-size-buffers=2 ! imxcompositor_g2d name=m
ix sink_0::zorder=2 sink_1::zorder=1 ! waylandsink
[  343.585808] enter isp_mi_stop
```

デモ①NNSTREAMER(NNSHARK)



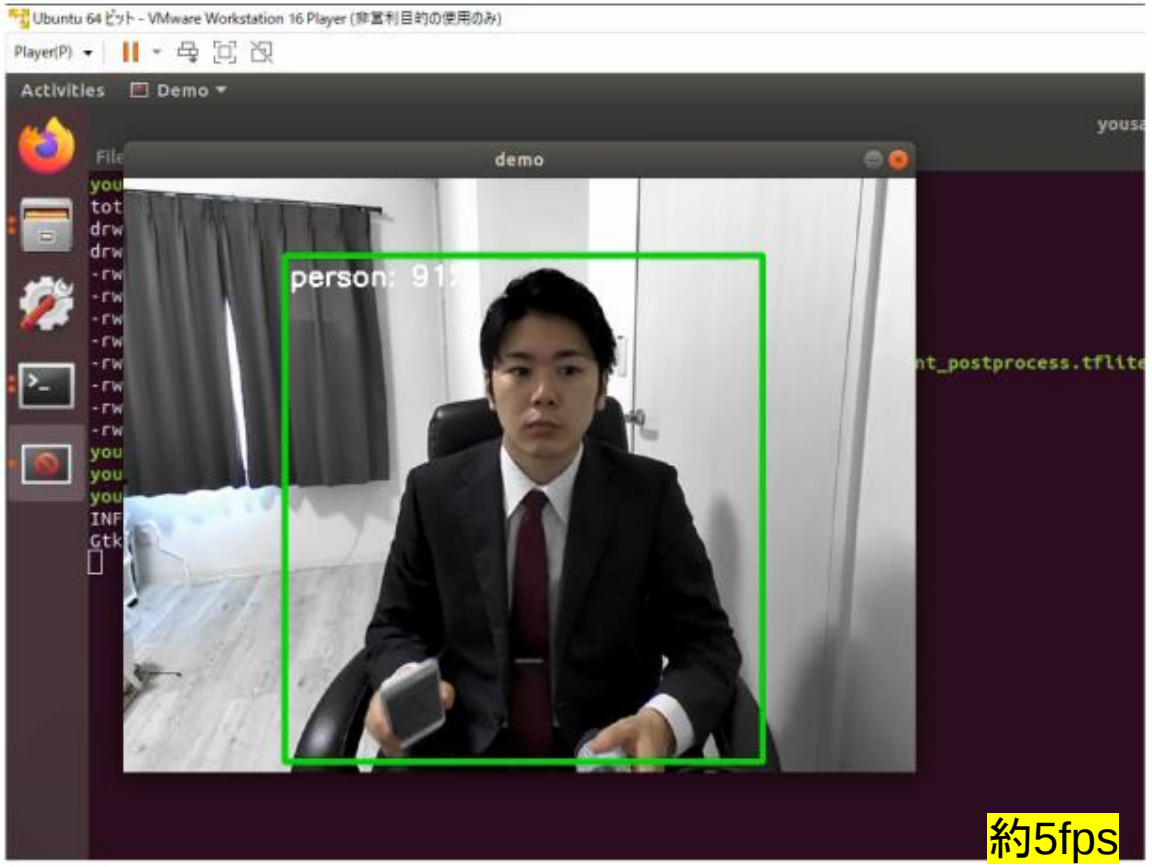
デモ①NNSTREAMER(NNSHARK)

↓i.MX 8M Plus Power Consumption Measurement抜粋

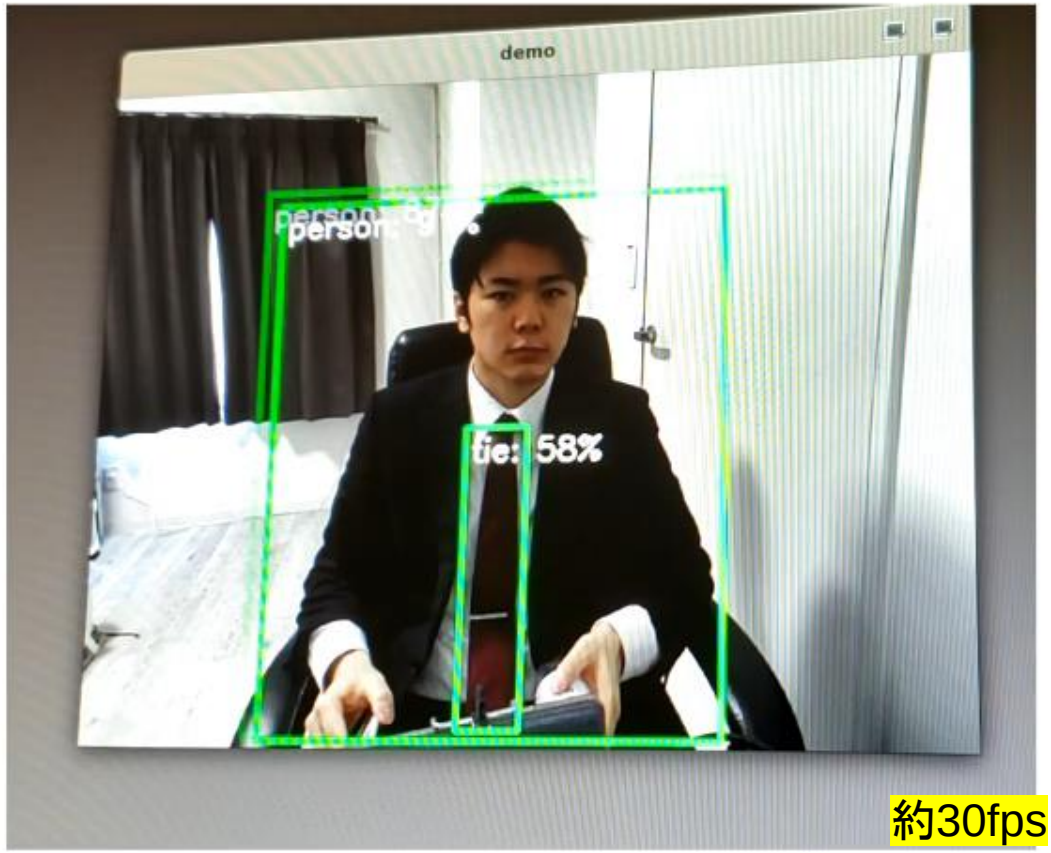
Power Group	Supply Domain	P(mW)	
		peak	avg
Group SOC	VDD_ARM	1004.20	65.37
	VDD_SOC	2576.67	1652.31
	NVCC_DRAM	376.83	174.54
	NVCC_SNVS_1P8	0.51	0.39
	VDD_xxx_1P81 + VDD_xxx _0P82 + VDD_USB_3P3	-	21.95
	Total	-	1914.57
Group DRAM	VDD1 + VDDQ + VDD2	-	259.26

デモ②処理性能比較

HostPC(x86@3.7GHz Dual core)にて実行



i.MX 8M Plus(NPU)にて推論実行

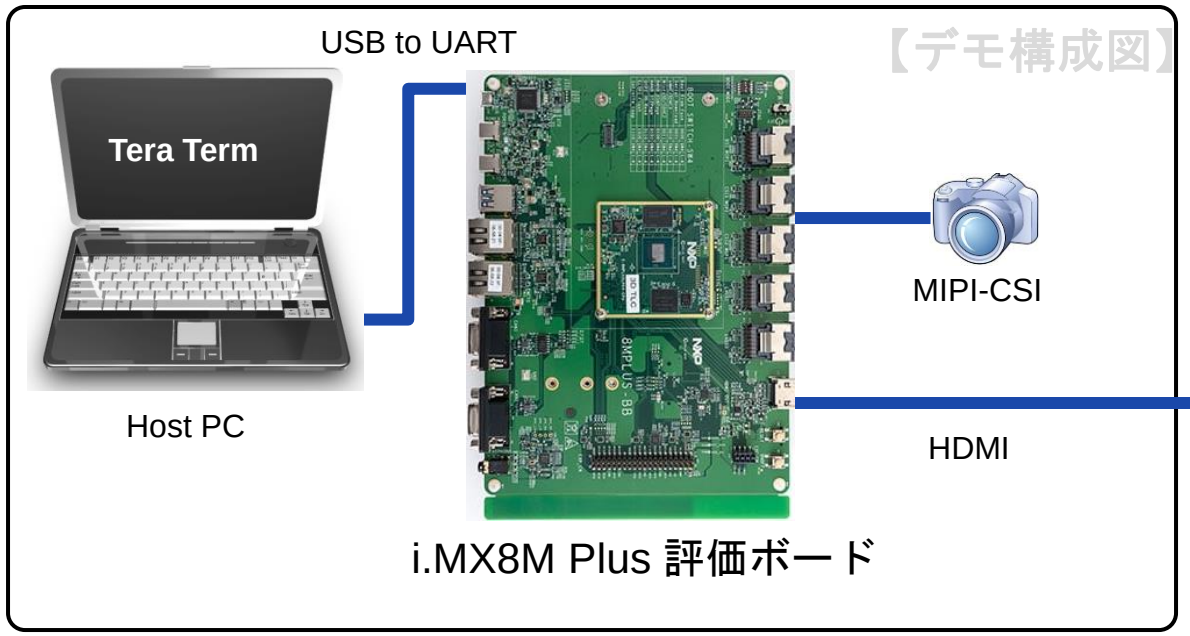


- 推論が完了するたびに画面をリフレッシュ
- 使用するモデル、スクリプト、解像度は同一

I.MX8M PLUSによるAIソリューションデモ

Forum会場で
デモ展示中

- NXPのi.MX8M Plus評価ボードとNXP提供の機械学習開発環境(eIQ)を使用した
  『Edgeデバイスによる推論デモ』を実施します



【デモの種類】
- **Object Classification：画像分類**
- **Object Detection：物体検出**
- **Pose Detection：姿勢検出**

5.まとめ

SECURE CONNECTIONS
FOR A SMARTER WORLD

まとめ

エッジAI向けOSS動向

- 様々なフレームワークがあるが、TensorFlow, Pytorchの人気が特に高い
- マルチメディア系OSSでDeep Learningを扱いやすい環境が整えられつつある

エッジAI導入におけるポイント

- Hardwareリソースが限定的
  ⇒モデル精度と処理速度のトレードオフを考慮する
  ⇒モデルを最適化（量子化/枝刈り）する

NXPのAI開発への取り組み(eIQ)

- eIQ Portalでモデルのトレーニング/評価/最適化などをGUIで実施可能
- i.MX 8M PlusはAI処理に特化したNPU(2.3TOPS)を搭載
- BSPに統合されているデモソフトで簡単かつ直ぐにエッジAIの評価を開始

SECURE CONNECTIONS
FOR A SMARTER WORLD