

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

PUBLIC

NXP, THE NXP LOGO AND NXP SECURE CONNECTIONS FOR A SMARTER WORLD ARE TRADEMARKS OF NXP B.V.
ALL OTHER PRODUCT OR SERVICE NAMES ARE THE PROPERTY OF THEIR RESPECTIVE OWNERS. © 2020 NXP B.V.



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

PUBLIC

NXP, THE NXP LOGO AND NXP SECURE CONNECTIONS FOR A SMARTER WORLD ARE TRADEMARKS OF NXP B.V.
ALL OTHER PRODUCT OR SERVICE NAMES ARE THE PROPERTY OF THEIR RESPECTIVE OWNERS. © 2020 NXP B.V.



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



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

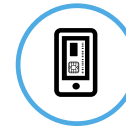
S32車載プラットフォーム
スケーラブル、セキュア、OTA、
ASIL D

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

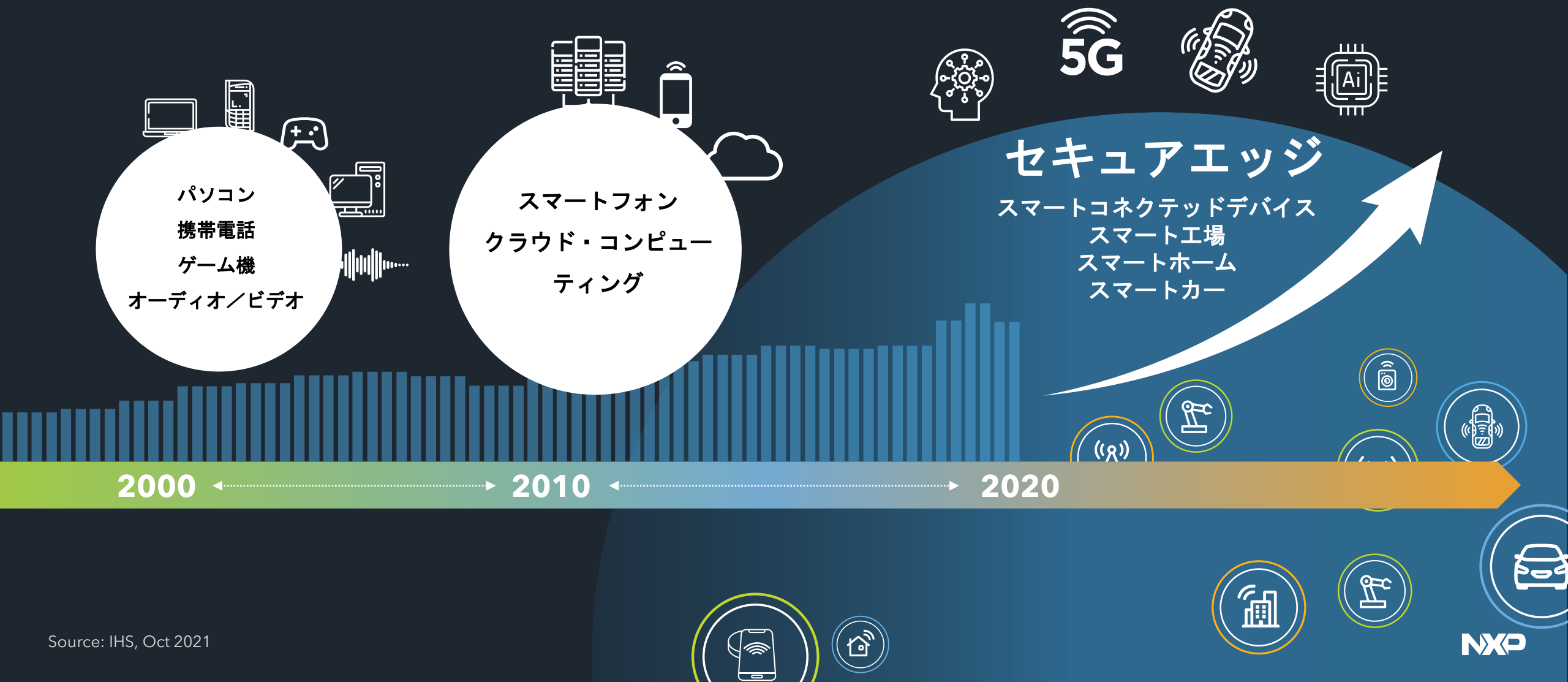
PUBLIC

NXP, THE NXP LOGO AND NXP SECURE CONNECTIONS FOR A SMARTER WORLD ARE TRADEMARKS OF NXP B.V.
ALL OTHER PRODUCT OR SERVICE NAMES ARE THE PROPERTY OF THEIR RESPECTIVE OWNERS. © 2020 NXP B.V.

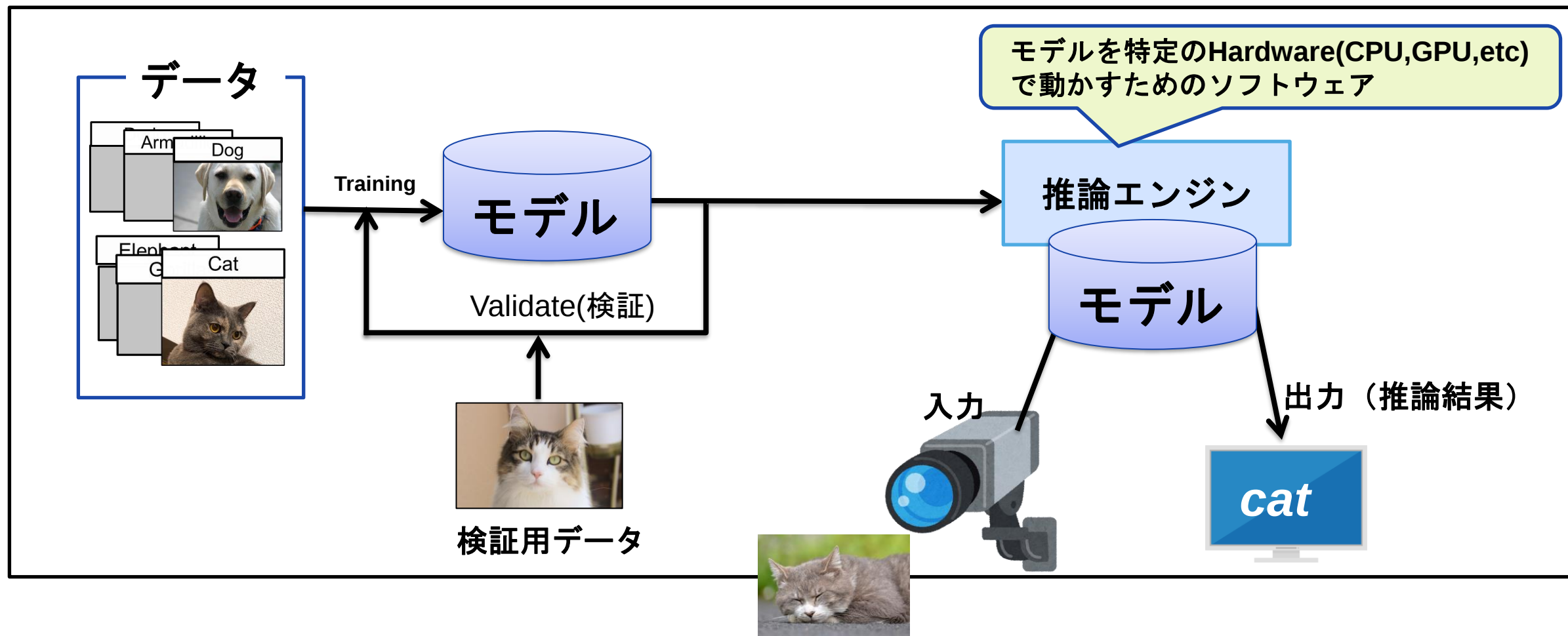


セキュアエッジの増加

2030年までにコネクテッド・デバイスが
750億台以上

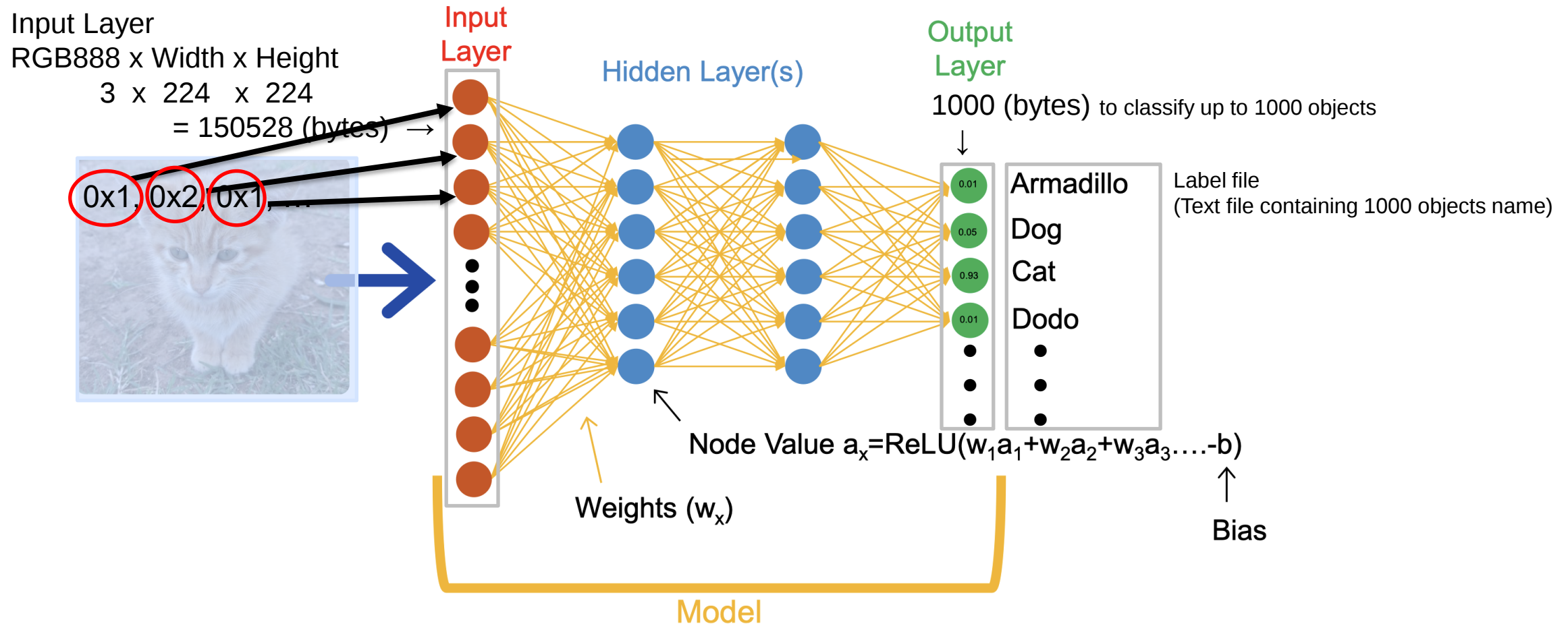


AI(Deep Learning)開発フロー 概要



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

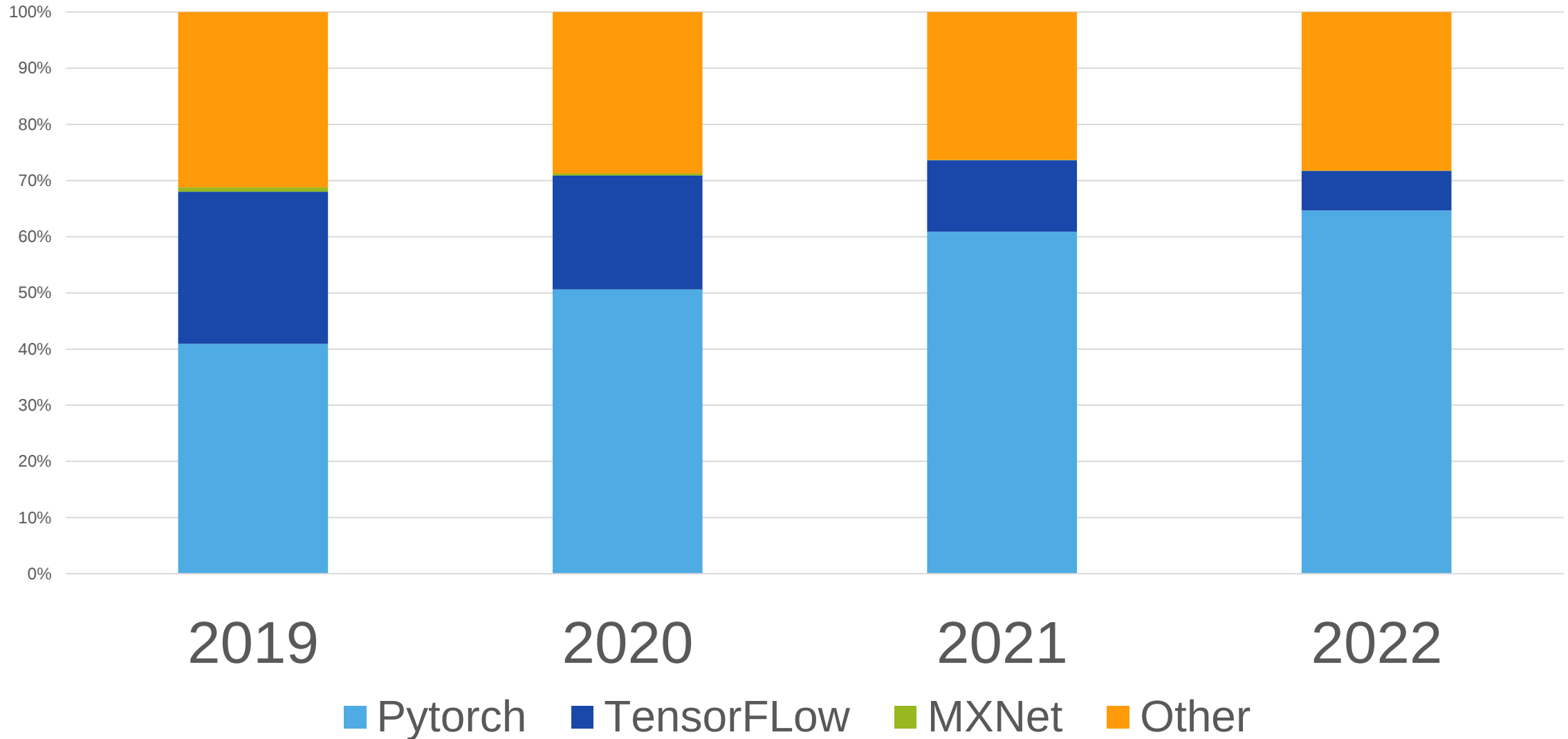
様々なDeep Learning Frameworkが存在
どれから勉強すればいいのか・・・

Deep Learning Frameworkの動向を調査！

人気があるOSSのDeep Learning Framework

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

論文でのDeep Learning Framework使用比率



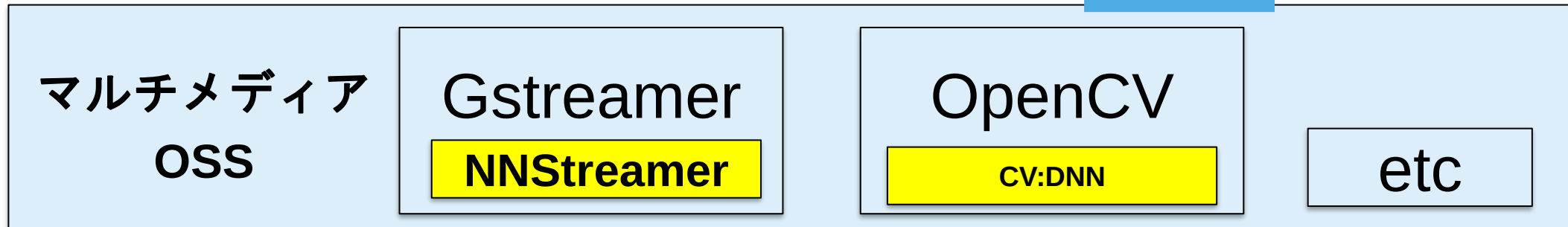
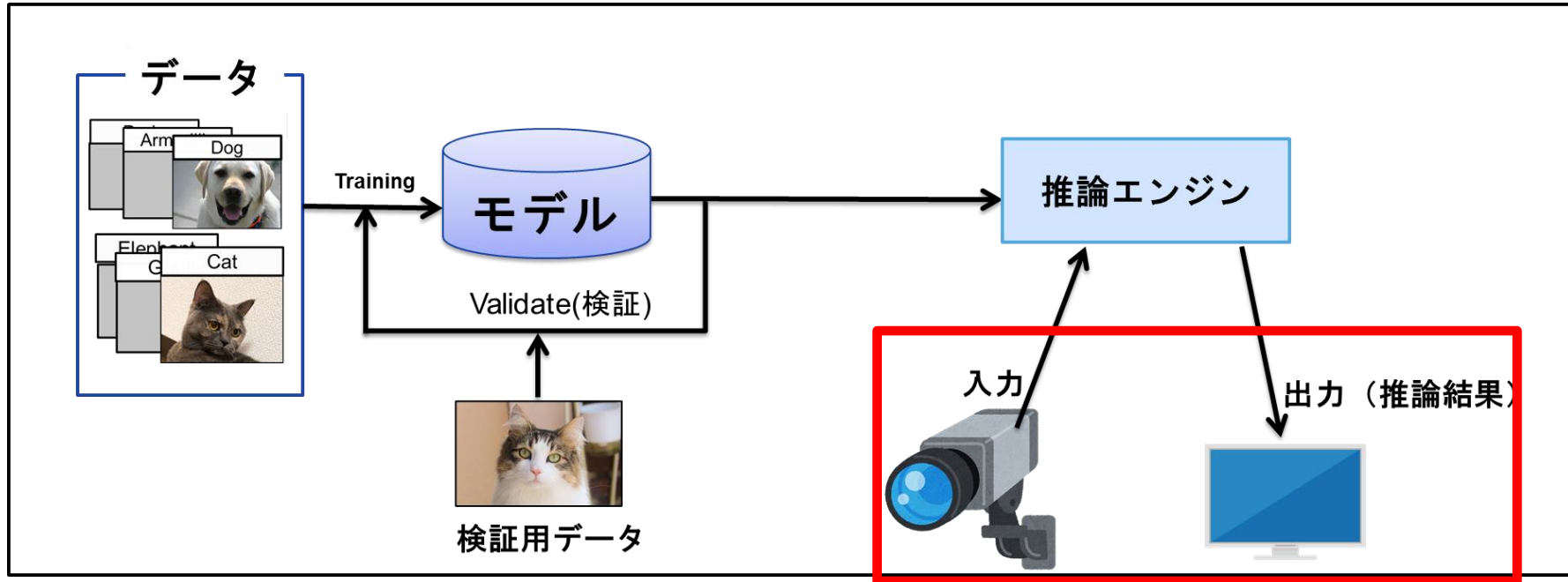
Define-By-RunとDefine-And-Run

	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

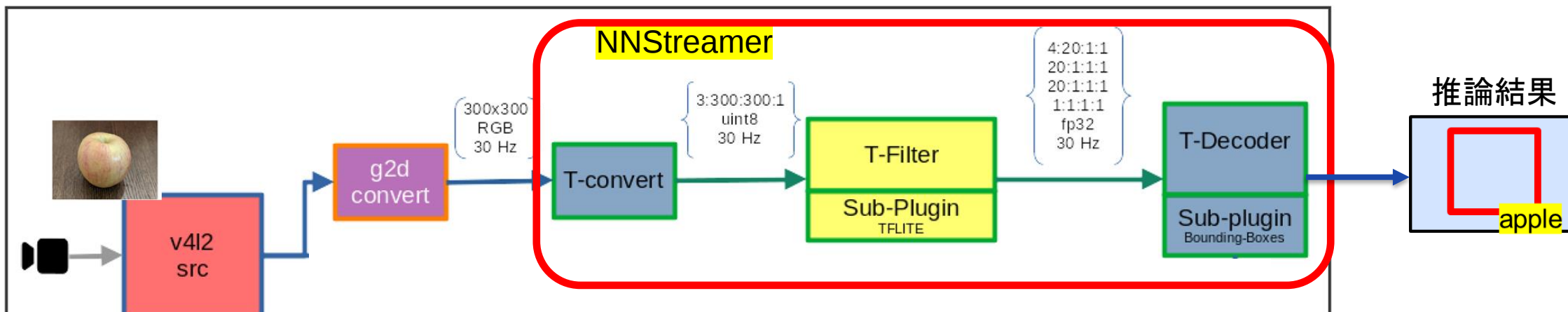
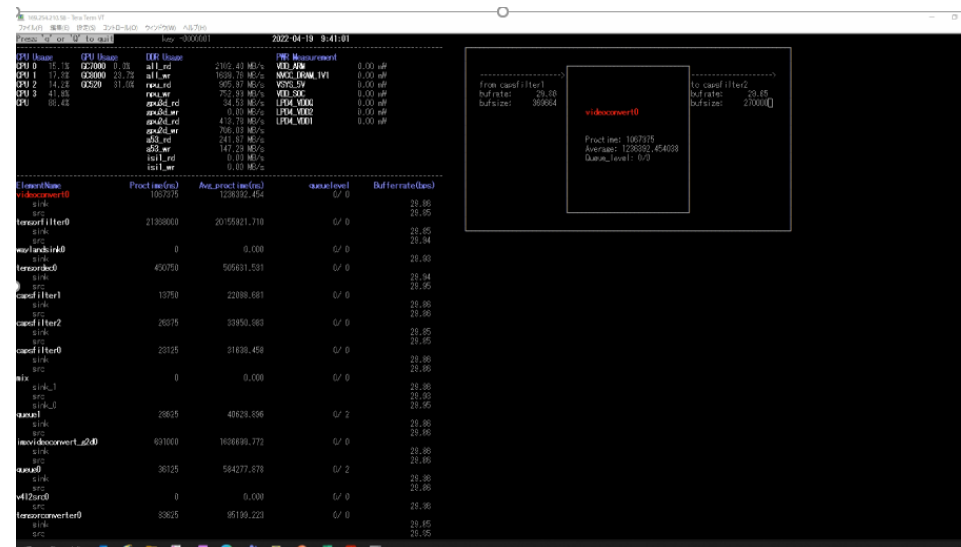
名称	開発元	対応言語	特徴
TensorFlow	Google	Python, C++ Java, Go	利用者が多い、Githubでも人気が高い Keras(高レベルAPI) で実装しやすい
Pytorch	Facebook	Python	研究分野での人気が高い Define-by-Run、デバッグしやすい
Apache MXNet	ワシントン大学, カーネギーメロン大学	Python,C++ Scala,R Matlab, Julia, etc	Define-and-Run 対応言語多 AWSが公式サポートを表明

DLアプリ(vision)開発で使われるOSS（マルチメディア系）



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

PUBLIC

NXP, THE NXP LOGO AND NXP SECURE CONNECTIONS FOR A SMARTER WORLD ARE TRADEMARKS OF NXP B.V.
ALL OTHER PRODUCT OR SERVICE NAMES ARE THE PROPERTY OF THEIR RESPECTIVE OWNERS. © 2020 NXP B.V.



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

入力画像サイズ
↓
MobileNet_v1_1.0_224

ハイパーパラメータ
1.0 : 精度最優先
0.25 : 処理速度最優先

「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

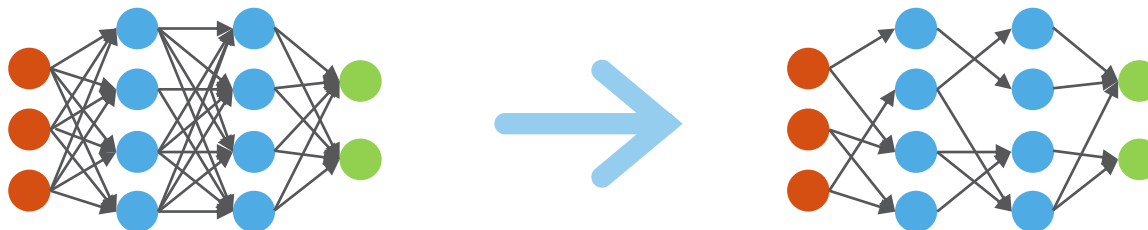
- モデルサイズを1/4に削減
- 整数処理は浮動小数点処理よりも高速
- 適切な処理で行えばaccuracy(精度)のロスが少ない



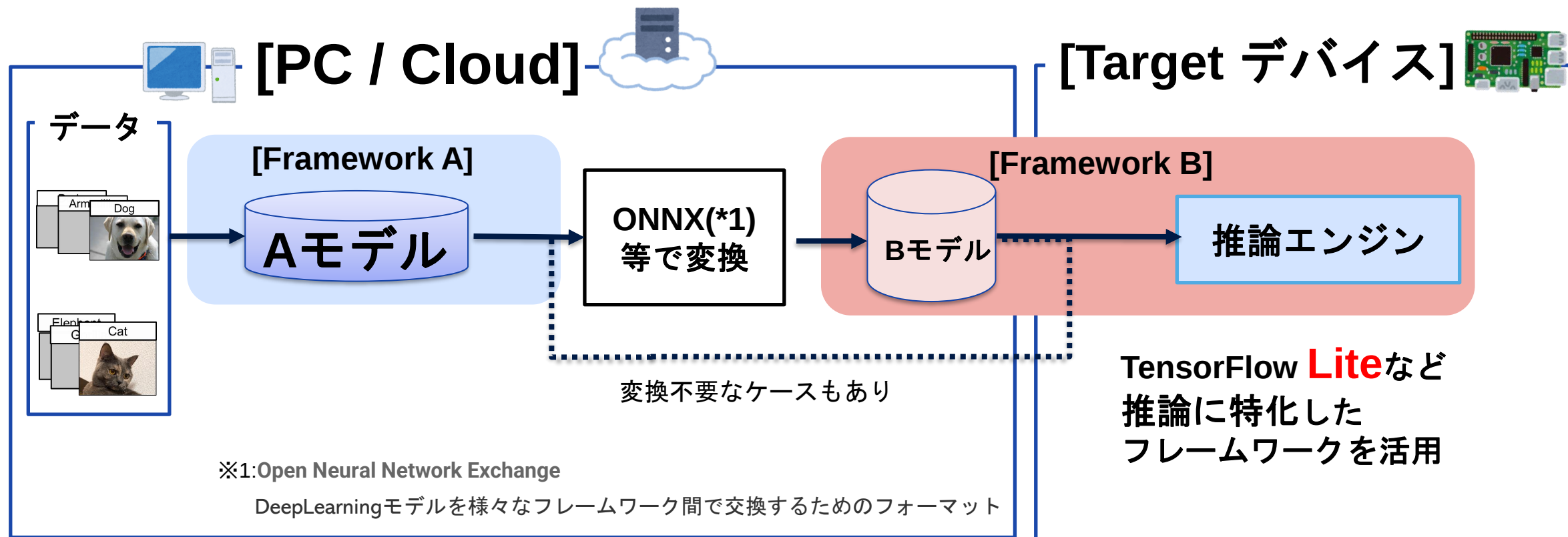
Pruning (枝刈り)

- 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

PUBLIC

NXP, THE NXP LOGO AND NXP SECURE CONNECTIONS FOR A SMARTER WORLD ARE TRADEMARKS OF NXP B.V.
ALL OTHER PRODUCT OR SERVICE NAMES ARE THE PROPERTY OF THEIR RESPECTIVE OWNERS. © 2020 NXP B.V.



eIQ機械学習ソフトウェア開発環境とは

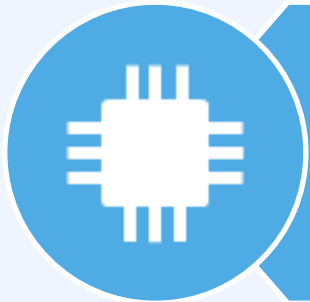
NXP製MCU/MPU向けMachine Learning(機械学習)アプリを構築するためのソフトウェアライブラリおよび開発ツール一式

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



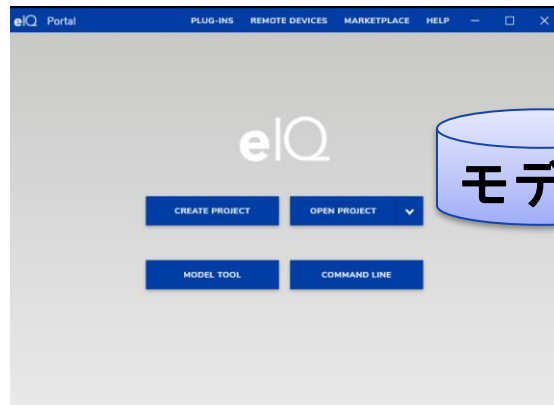
- ・モデルトレーニングツール
- ・無償利用可能

eIQ 機械学習ソフトウェア開発環境

eIQ



[Host PCで動作]
eIQ Portal



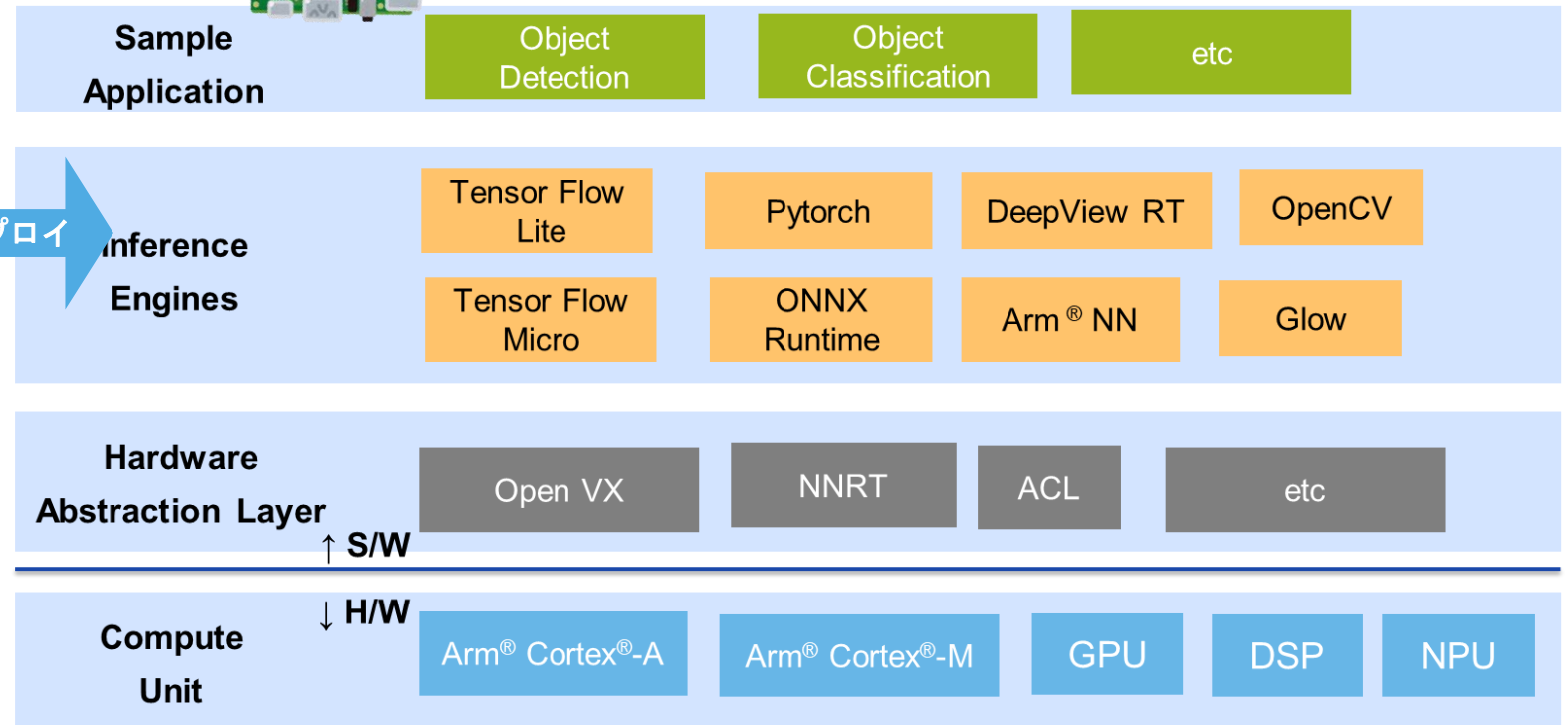
モデル

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)



eIQ Portal紹介(データ管理)

eIQ Portal flower

PLUG-INS REMOTE DEVICES WORKSPACES MARKETPLACE HELP

Labels	Number of Images		
	Train	Test	Total
daisy	521	56	577
dandelion	786	56	842
roses	529	56	585
sunflowers	587	56	643
tulips	687	56	743

Unlabeled Images 0

All Images 3110 280 3390

AUGMENTATION TOOL

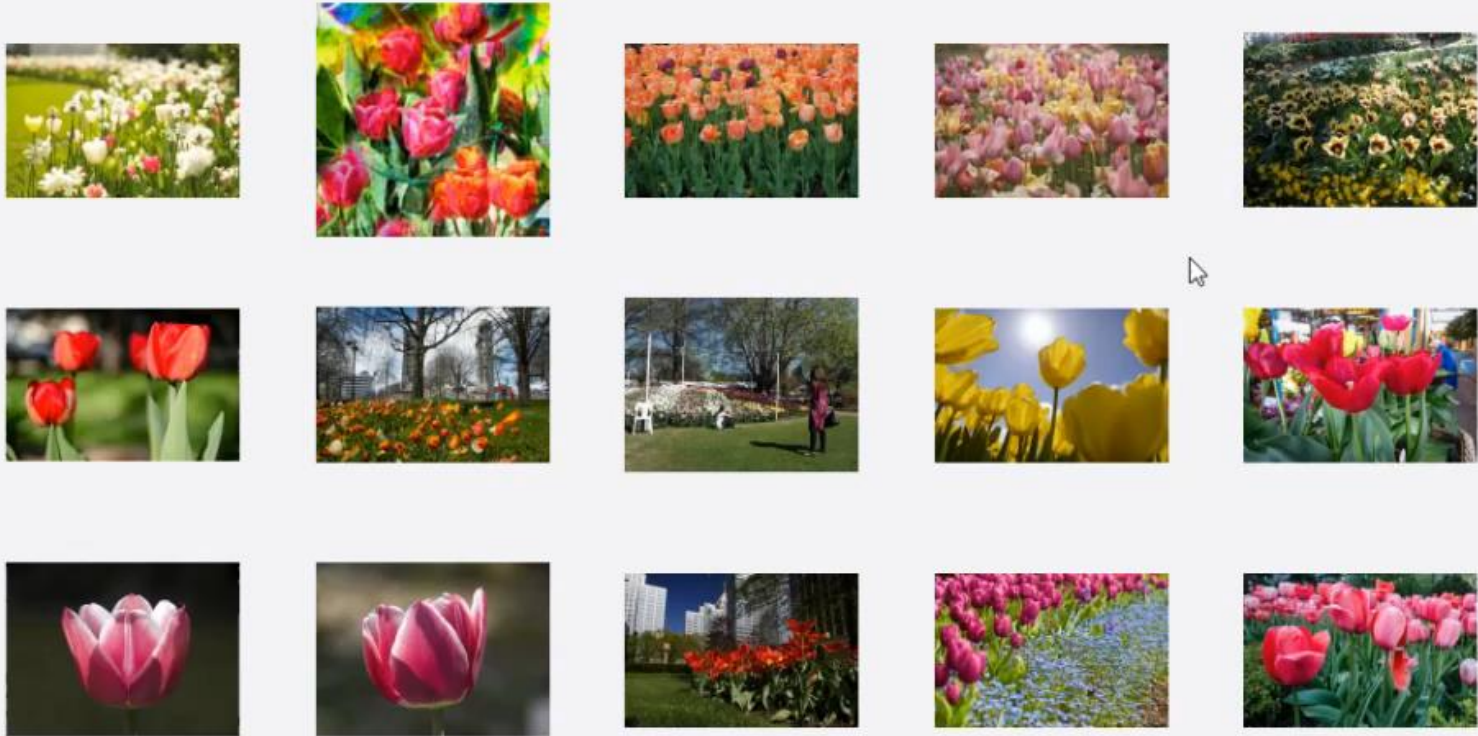
DEEVIEW DEVPACK ADD-ON

< HOME SELECT MODEL >

Data Set Curator

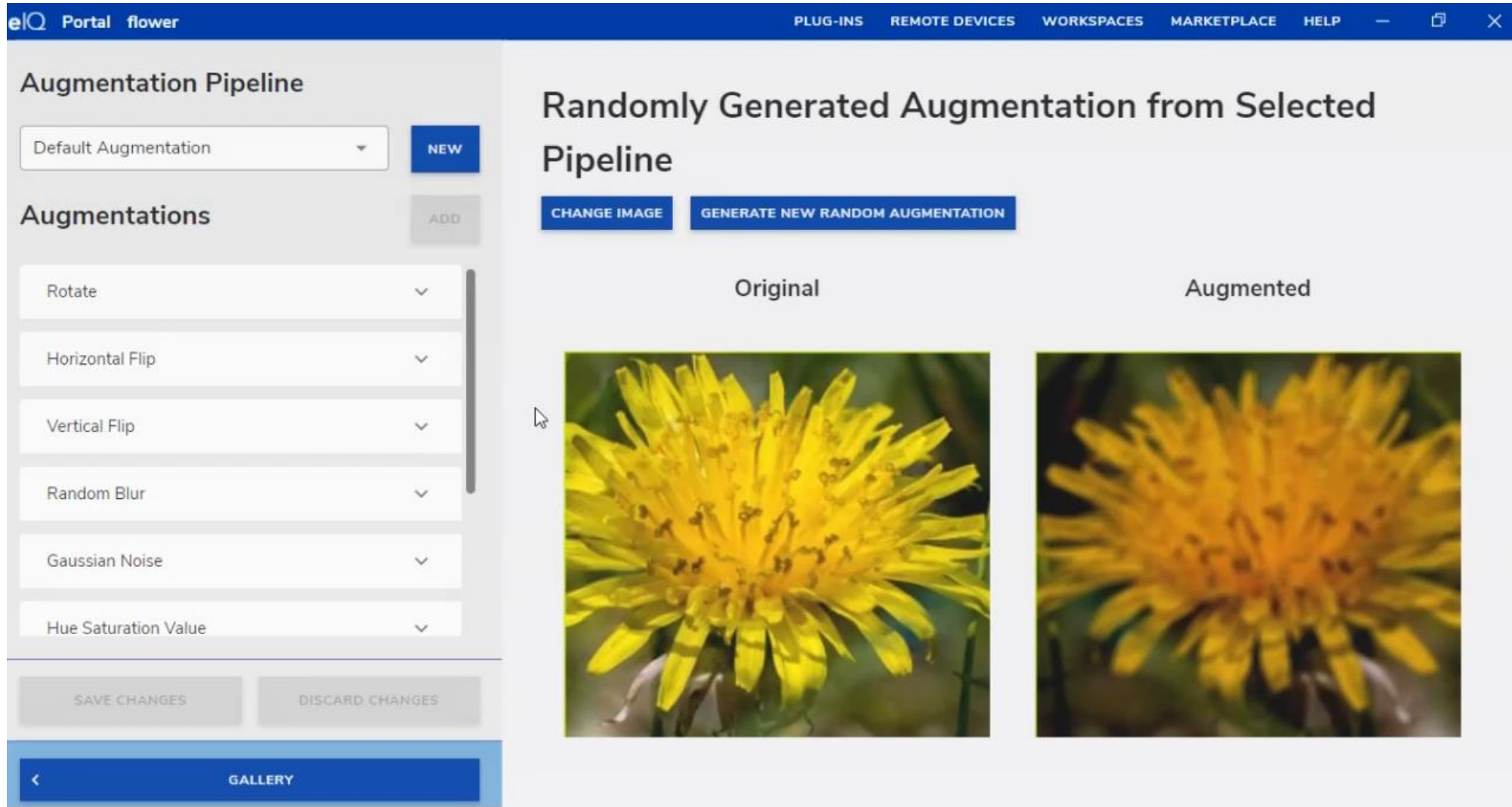
IMPORT CAPTURE

Dataset Test Holdout 20 % SHUFFLE



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)



Performance

The performance-optimized model seeks to achieve the lowest inference latency possible at the cost of accuracy. Choose this model if you require maximum performance or are using a resource-constrained microcontroller.



Balanced

The balanced model aims to have performance and accuracy results that strike a balance between the performance model and the accuracy model. Choose this model if you are seeking a balance of



Accuracy

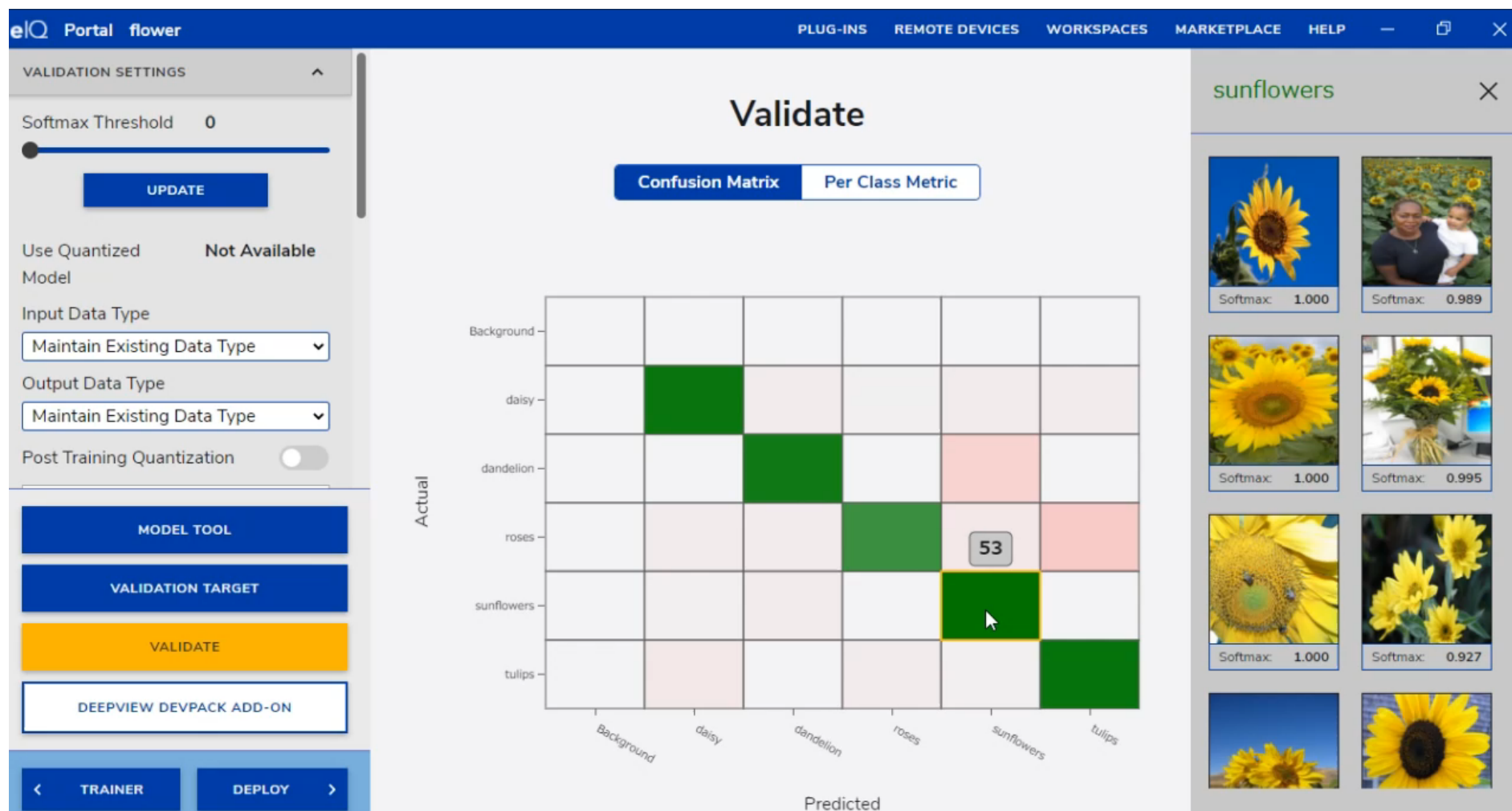
The accuracy-optimized model aims to provide the best possible accuracy with reasonable performance on a typical embedded platform. Choose this model if you require the best possible accuracy and are less sensitive to inference latency.

**「速度」「精度」の
優先度を選択可能**

eIQ Portal紹介(トレーニング)

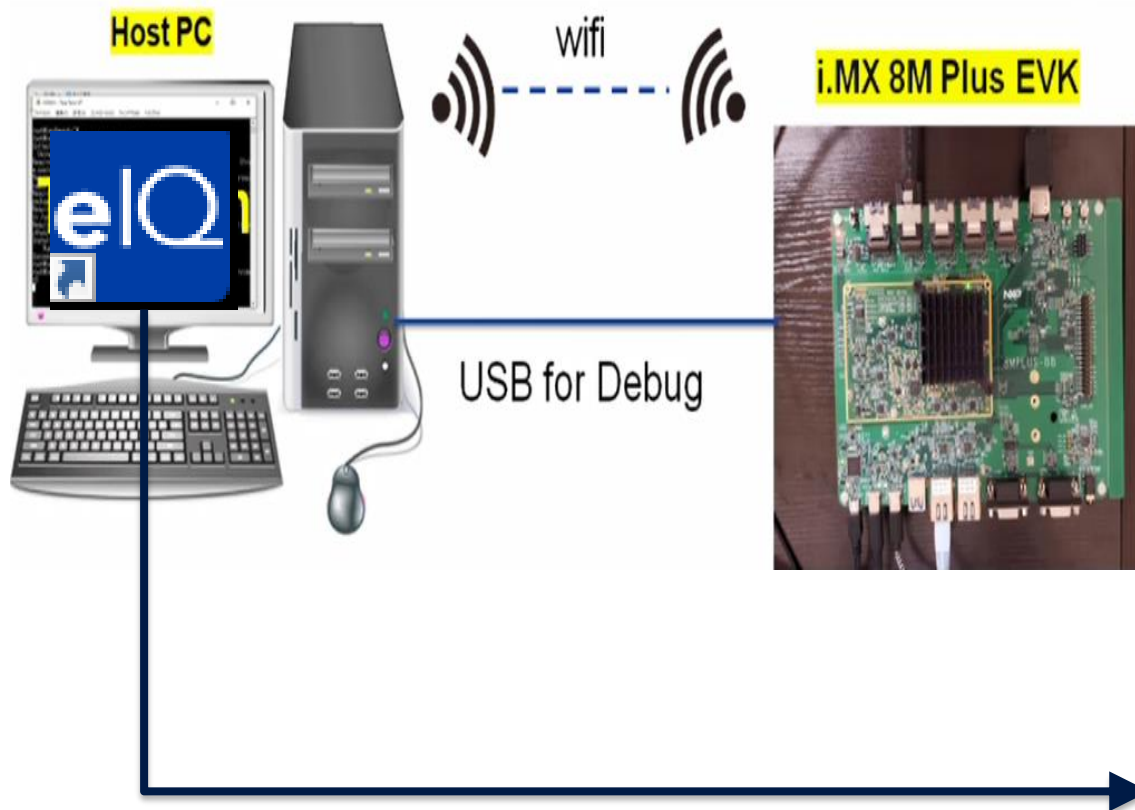


eIQ Portal紹介(検証)



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)

The screenshot shows the eIQ Portal interface for exporting a model. On the left, two callout boxes highlight specific settings: the 'Export File Type' dropdown is set to 'Deepview RT', and the 'Input Data Type' dropdown is set to 'Maintain Existing Data Type'. In the main 'EXPORT SETTINGS' panel, the 'Export Quantized Model' toggle is turned on, and the 'Per Channel' option is selected. A red box highlights the 'Number of Samples' field, which is set to 10. A yellow callout box with the text '←モデルの最適化（量子化）' points to the quantization settings. The right panel, titled 'Export Model', displays the project details and training metrics.

Export Model	
Project	flower
Model Name	classification-precision-npu-2021-08-11T07-59-07.251Z
Labels	daisy, dandelion, roses, sunflowers, tulips
Epochs Trained	4
Training Time	1235s
Training Accuracy	94.8%
Validation Accuracy	85.36%

モデル変換機能

GUI操作でモデルの変換可

The screenshot displays the eIQ Model Tool 2.4.2 interface. The 'File' menu is open, showing options: Open..., Open Recent, Export..., Convert, Close, and Exit. The 'Convert' option is highlighted, and a submenu is visible with the following options: Deepview RT (.rtm), ONNX (.onnx), and Tensorflow Lite (.tflite). The main workspace shows a neural network diagram with the following components and data flow:

- Input_1**: Initial input node.
- Quantize**: Operation node receiving input from Input_1. Data flow: $1 \times 320 \times 320 \times 3$.
- Conv2D**: Operation node receiving input from Quantize. Data flow: $1 \times 320 \times 320 \times 3$. Parameters: filter $(16 \times 3 \times 3 \times 3)$, bias (16) .
- Add**: Operation node receiving input from Conv2D. Data flow: $1 \times 160 \times 160 \times 16$. Parameters: B = 127, Relu6.
- Mul**: Operation node receiving input from Conv2D and Add. Data flow: $1 \times 160 \times 160 \times 16$.
- DepthwiseConv2D**: Operation node receiving input from Mul. Data flow: $1 \times 160 \times 160 \times 16$. Parameters: weights $(1 \times 3 \times 3 \times 16)$, bias (16) , Relu.

i.MXプロセッサ +eIQによるデモ



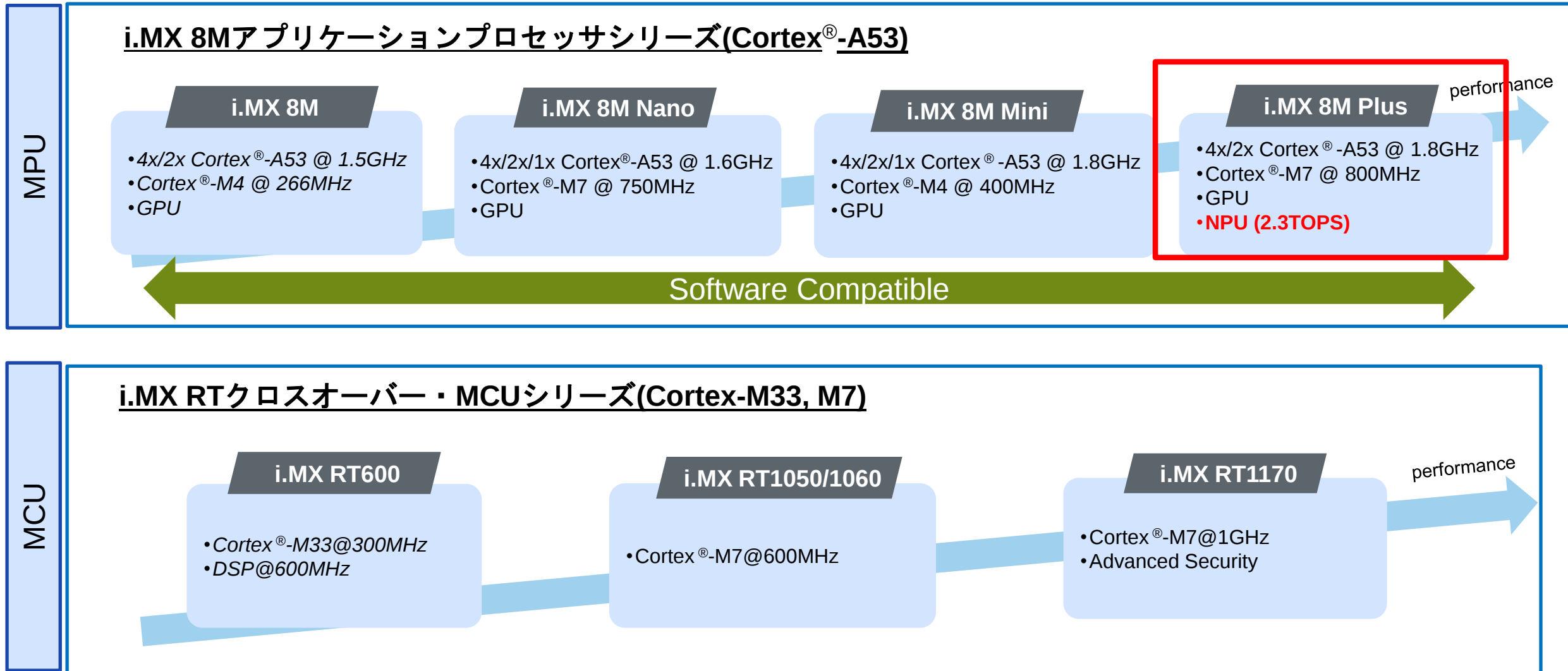
SECURE CONNECTIONS
FOR A SMARTER WORLD

PUBLIC

NXP, THE NXP LOGO AND NXP SECURE CONNECTIONS FOR A SMARTER WORLD ARE TRADEMARKS OF NXP B.V.
ALL OTHER PRODUCT OR SERVICE NAMES ARE THE PROPERTY OF THEIR RESPECTIVE OWNERS. © 2020 NXP B.V.



eIQ対応MPU/MCUラインナップ



デモ①NNSTREAMERを用いたOBJECT DETECTION



デモ①NNSTREAMER(NNSHARK)

192.168.0.181 - 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)

Press 'q' or 'Q' to quit key -0000001 2021-11-19 17:32:10

CPU Usage	GPU Usage	DDR Usage	PWR Measurement
CPU 0 10.8%	GC7000 nan%	all_rd 2047.82 MB/s	VDD_ARM 232.10 mW
CPU 1 22.2%	GC8000 nan%	all_wr 1672.56 MB/s	NVCC_DRAM_1V1 178.00 mW
CPU 2 14.1%	GC520 nan%	npv_rd 876.38 MB/s	VSYS_5V 2982.00 mW
capsfilter2%		npv_wr 763.46 MB/s	VDD_SOC 1248.00 mW
frate: 30.51		gpu3d_rd 0.00 MB/s	LPD4_VDDQ 30.77 mW
fsize: 270000		gpu3d_wr 0.00 MB/s	LPD4_VDD2 359.80 mW
		gpu2d_rd 391.03 MB/s	LPD4_VDD1 9.57 mW
		gpu2d_wr 703.50 MB/s	
		a53_rd 282.36 MB/s	
		a53_wr 170.77 MB/s	
		isi1_rd 0.00 MB/s	
		isi1_wr 0.00 MB/s	

ElementName	Proctime(ns)	Avg_proctime(ns)	queuelevel	Bufferrate(bps)
videoconvert0	1054500	1229217.051	0/ 0	
sink				30.49
src				30.51
tensorfilter0	19817875	20254905.256	0/ 0	
sink				30.52
src				30.51
waylandsink0	0	0.000	0/ 0	
sink				30.04
tensordec0	649250	557148.822	0/ 0	
sink				30.51
src				30.51
capsfilter1	12750	23625.897	0/ 0	
sink				30.49
src				30.49
capsfilter2	27875	34797.821	0/ 0	
sink				30.51
src				30.51
capsfilter0	35375	32768.018	0/ 0	
sink				29.70
src				29.70

```

----->
from capsfilter1
bufrate: 30.49
bufsize: 369664

videoconvert0

Proctime: 1054500
Average: 1229217.051282
Queue_level: 0/0
--
to bu
bu

```

I

デモ①NNSTREAMER(NNSHARK)

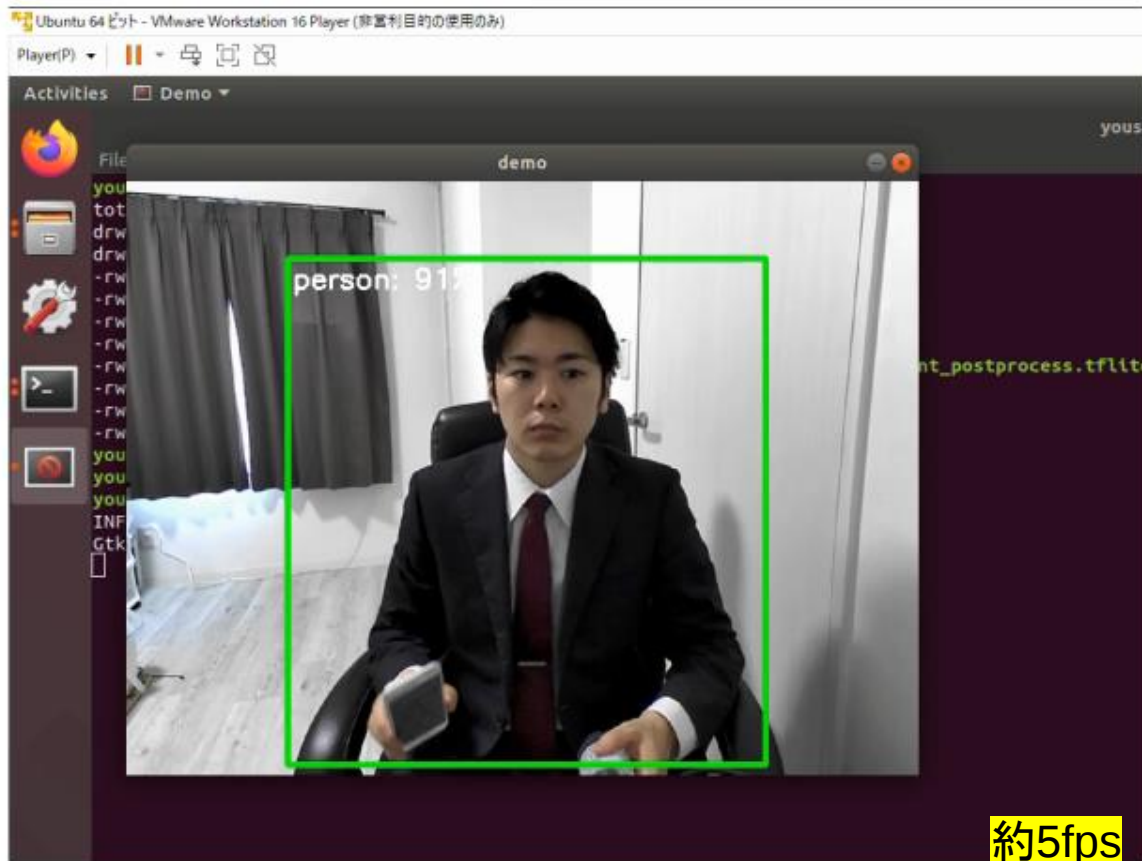
↓i.MX 8M Plus Power Consumption Measurement抜粋

PWR Measurement		
VDD_ARM	74.10	mW
NVCC_DRAM_1V1	223.10	mW
VSYS_5V	2240.00	mW
VDD_SOC	1137.00	mW
LPD4_VDDQ	21.90	mW
LPD4_VDD2	125.70	mW
LPD4_VDD1	9.59	mW

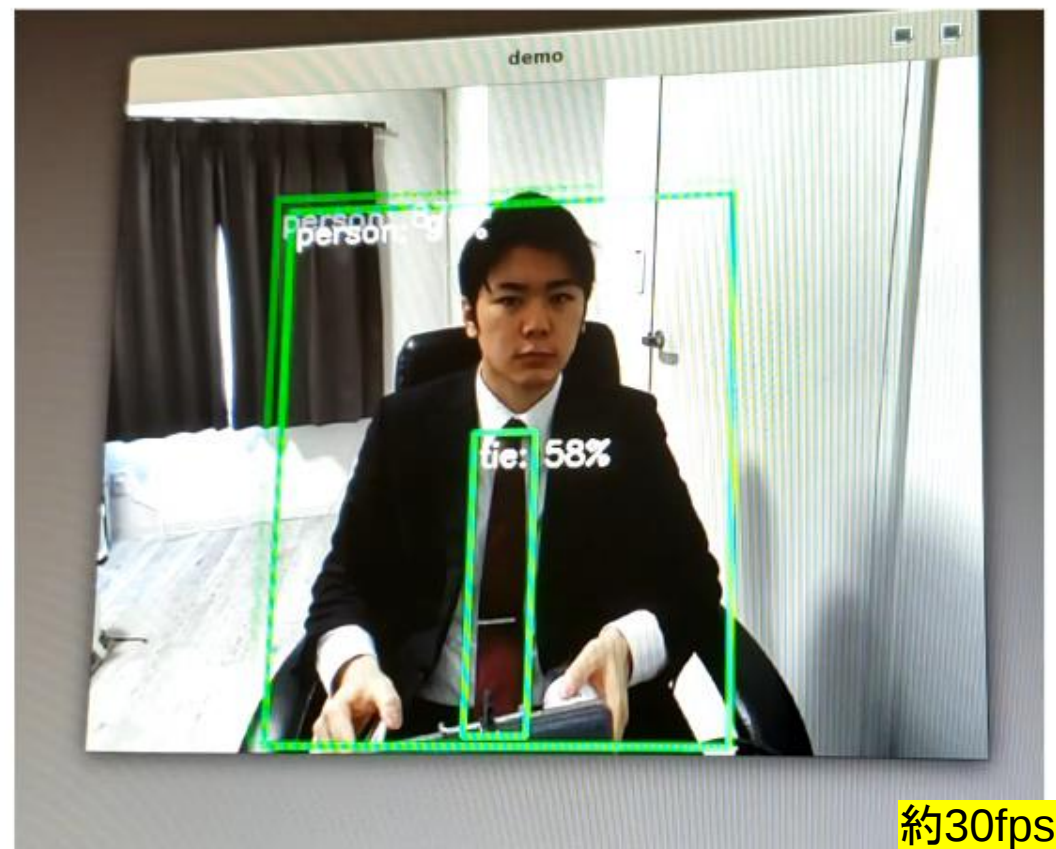
Power Group	Supply Domain	P(mW)	
		peak	avg
Group SOC	VDD_ARM	1004.20	65.37
	VDD_SOC	2576.67	1652.31
	NVCC_DRAM	376.83	174.54
	NVCC_SNVIS_1P8	0.51	0.39
	VDD_1P81 + VDD_0P82 + VDD_USB_3P3	-	21.95
	Total	-	1914.57
Group DRAM	VDD1 + VDDQ + VDD2	-	259.26

デモ②処理性能比較

HostPC(x86@3.7GHz Dual core)にて実行



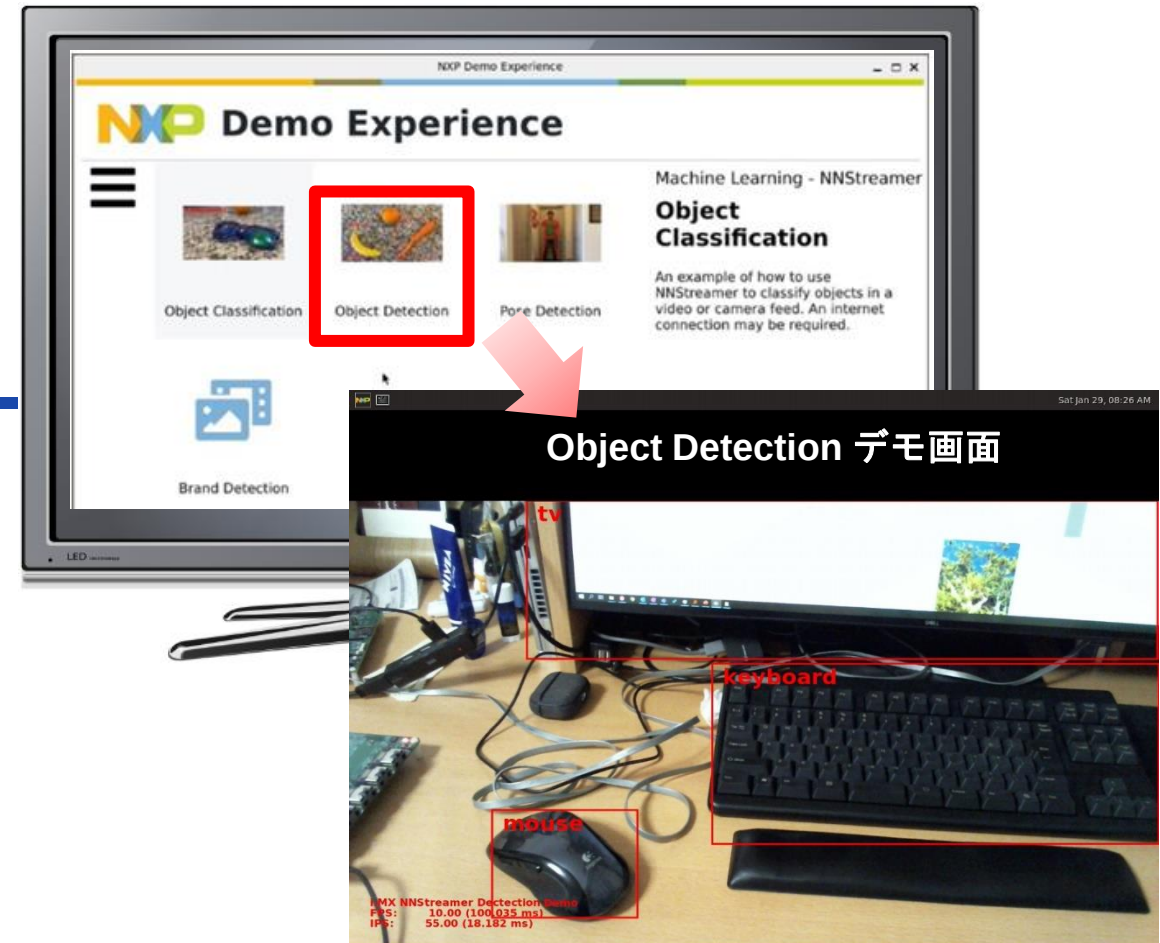
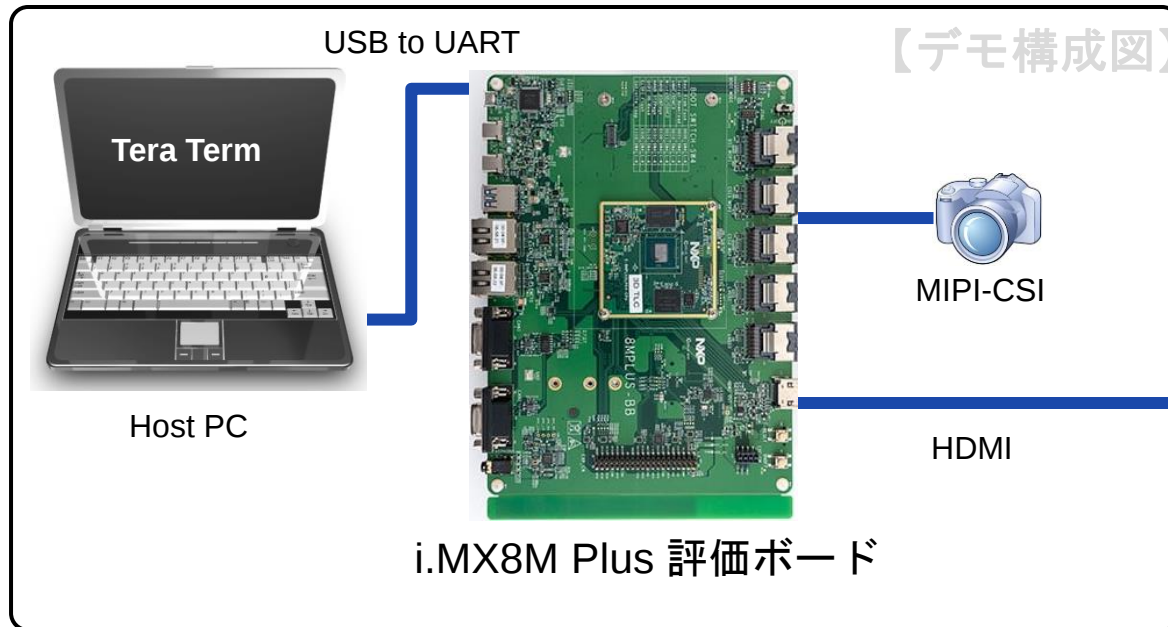
i.MX 8M Plus(NPU)にて推論実行



- ・ 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

- 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

PUBLIC

NXP, THE NXP LOGO AND NXP SECURE CONNECTIONS FOR A SMARTER WORLD ARE TRADEMARKS OF NXP B.V.
ALL OTHER PRODUCT OR SERVICE NAMES ARE THE PROPERTY OF THEIR RESPECTIVE OWNERS. © 2020 NXP B.V.



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の人気の特に高い
- 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