MATLAB EXPO 2018

ディープラーニングの
システムへの展開
～エッジからクラウドまで～

アプリケーションエンジニアリング部
福本　拓司

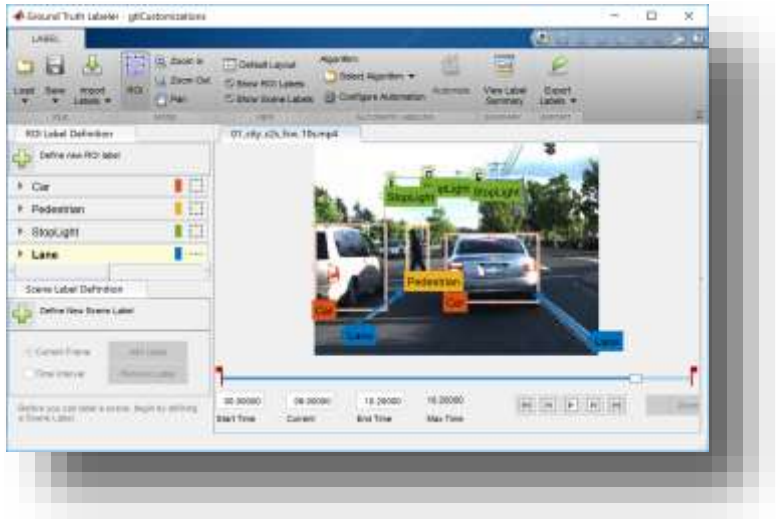
機械学習・ディープラーニング関連セッション

エンジニアリング データアナリティクス	ディープラーニング	生産技術	自動運転	モーター制御と電力制御	ロボティクス	アドバンスト
A1 脳情報通信技術の理論と実践 ～未来を切り拓く多様で多次元な「人間」に関わるアナリティクス 株式会社ＮＴＴデータ経営研究所 茨木 拓也	B1 ディープラーニングを活用した山岳トンネルの岩盤評価 株式会社大林組 畑 浩二	C1 人工知能の眼による検品自動化 ～生産現場で価値を生むAIシステム～ 武蔵精密工業株式会社 神谷 文久	D1市街地における自動車の自動運転に向けて 金沢大学 新学術創成研究機構 未来社会創造研究コア 菅沼 直樹	E1 モーター制御開発のMBDトレーニングとバッテリー充放電コントローラの機能安全対応事例 パナソニック アドバンストテクノロジー株式会社 堀江 雅浩	F1 MBDを活用した福島第一原子力発電所燃料デブリ取出しロボットの設計開発 三菱重工業株式会社 村田 直史	G1 競技用自転車開発における挙動推定と1Dシミュレーションの活用 株式会社ブリヂストン 内田 和男
休憩						
A2 機械学習を1週間でビジネスに活用させる3つの方法 -故障予知を事例に 大阪ガス株式会社 國政 秀太郎	B2 ディープラーニングの実践的な適用ワークフロー *MathWorks Japan* 蘇 亮	C2 外観検査のための画像処理・ディープラーニングワークフロー *MathWorks Japan* 町田 和也	D2 ADSTを使用した合流・交差点シーンにおけるADの行動解析 株式会社本田技術研究所 上田 雄悟 *MathWorks Japan* 船田 龍平	E2 電力システムのモデリング・解析環境構築と工場受変電システムの解析 富士電機株式会社 壹岐 浩幸 九州工業大学 渡邊 政幸	F2 モデルベースデザインではじめる自律型アームロボットの開発・導入 *MathWorks Japan* 小林 昇洋	G2 設計・開発プロセス IEC61508 機能安全規格 SIL4 適合に向けた検証フローの構築 株式会社日立製作所 中村 昌平 上薗 巧
休憩						
A3 センサー/テキストデータ解析分野への機械学習の適用 *MathWorks Japan* 井原 瑞希	B3 大規模計測と高度な数理的処理の組み合わせが実現するインフラメンテナンスの革新的近未来 東京大学 水谷 司	C3 工程検査におけるディープラーニング活用による品質改善への取り組み デクセリアルズ株式会社 大河原 秀之 松下 忍	D3 ADAS・自動運転の開発・検証を加速する MATLAB/Simulink *MathWorks Japan* 大塚 慶太郎	E3 スマートグリッドなどの電力システム解析および制御への適用事例 芝浦工業大学 藤田 吾郎	F3 トヨタが開発するロボットHSRによる生活支援アプリケーション トヨタ自動車株式会社 池田 幸一 *MathWorks Japan* 木川田 亘	G3 無線モデム FPGA/SoC開発におけるHDL Coderの活用事例 *NEC*ネットワーク・センサ株式会社 住田 憲昭 *MathWorks Japan* 田中 明美

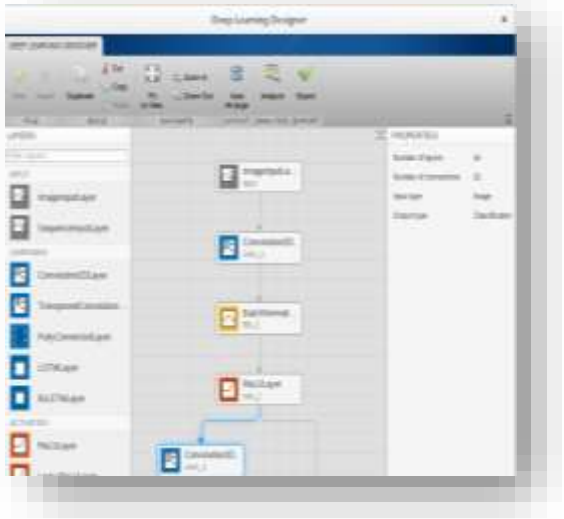
ディープラーニング学習のイメージできましたでしょうか？

・カメラ、データベースでのデータ取得
・簡潔なコーディングで学習&検証
・豊富なサンプルコード
・ユーザー成功事例

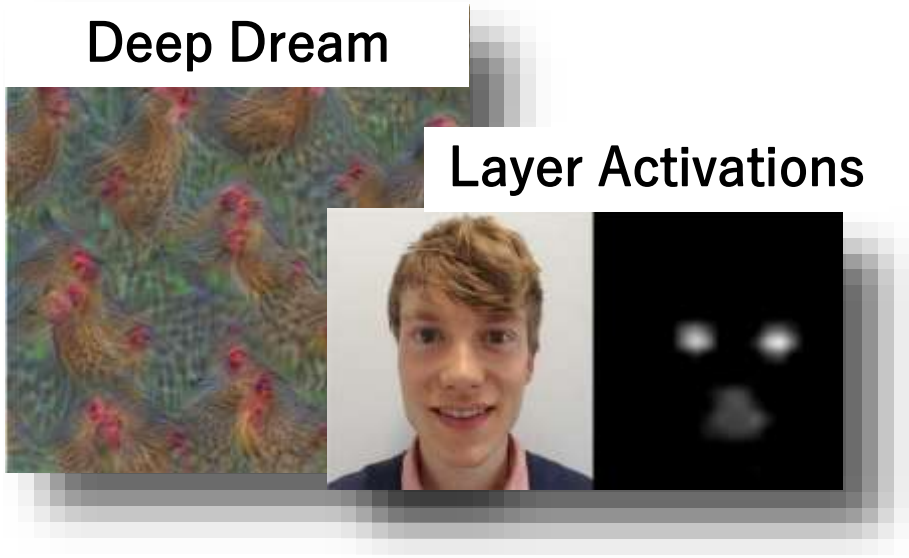
ラベリング

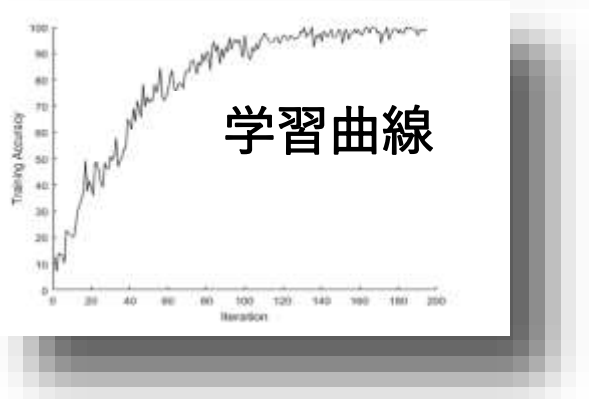


モデル作成



可視化





ネットワーク学習からシステムへの展開へ

マルチGPU　　ワークステーション　　クラウド

モデル学習環境

- パラメータ調整
  高精度を目指す

速度　・　コスト　・　使いやすさ

システム展開

高い精度だけでなく、　利用用途・環境に最適化された運用が必要

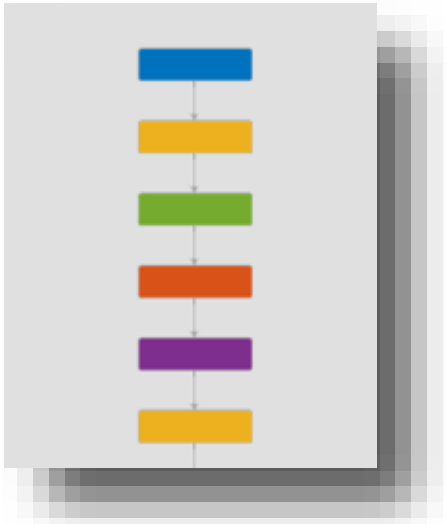
アジェンダ

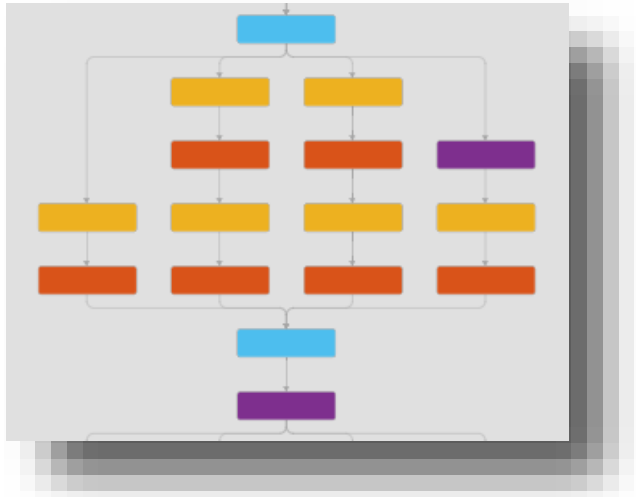
システム展開のためのMATLAB®活用術

システム化を踏まえたモデル選択

ターゲットの選択とシステム展開

ネットワークタイプの特徴





シリーズネットワーク		DAGネットワーク
大きい	ネットワークサイズ	小さい
演算がシンプル 高速	演算量に対する 推論速度	メモリアクセスが複雑 低速
AlexNet,　VGG他	モデル例	GoogleNet, ResNet他

推論の高速化 & メモリ消費の削減

「精度はばっちり。AlexNetは1000種類、私たちは数種類の分類。
もう少しサイズ小さくならないですか？」

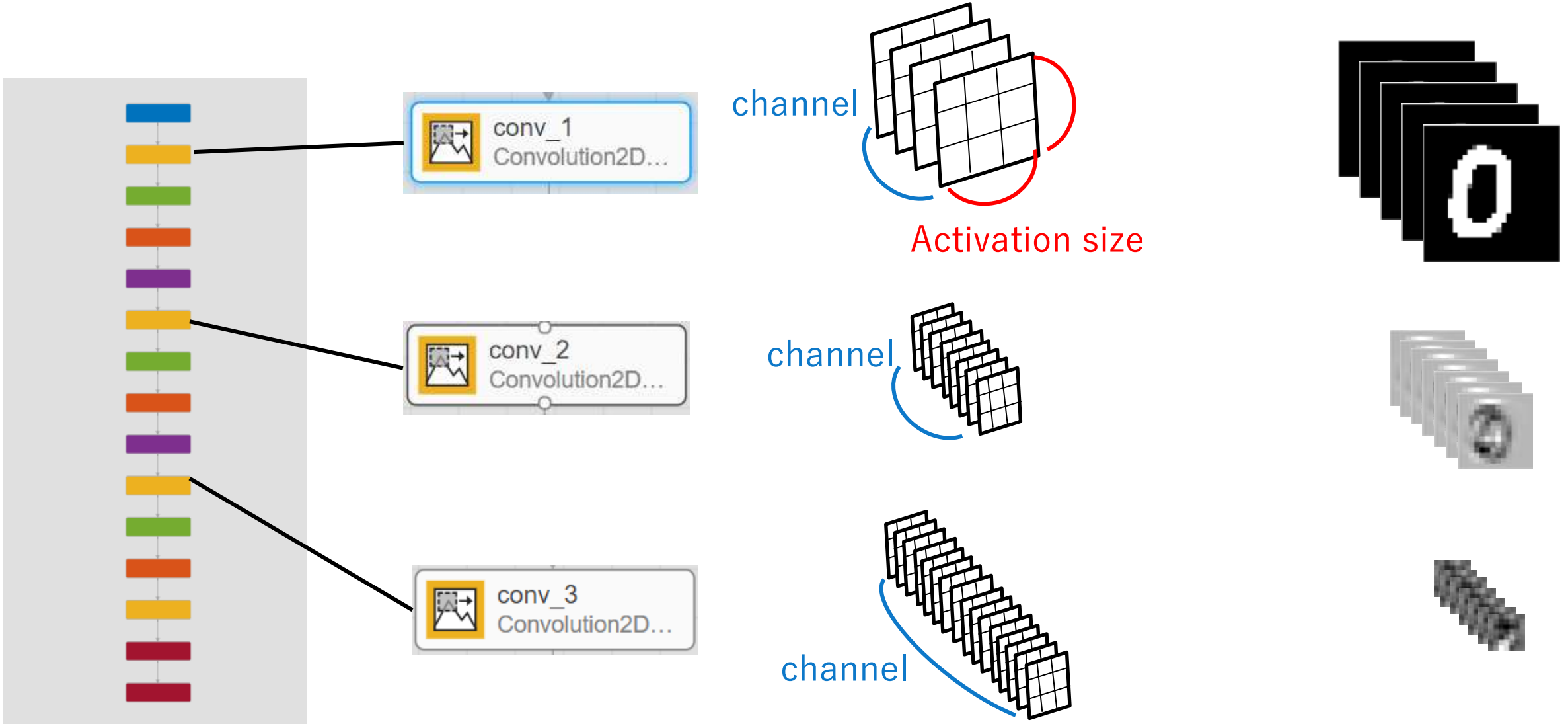
※ AlexNet　＞ 200MB

速度面を気にする → Shallowな層を組んでみましょう
GPU Coderを活用しましょう

ネットワークサイズを気にする → SqueezeNet(≒5MB)がお勧めです

再学習モデルを効率化したい → モデルを修正する方法があります。

学習済みネットワークの最適化



学習済みネットワークの最適化

取り除く
channel

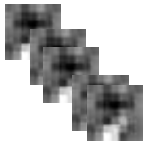
channel

Channelを減らせば演算量が減る
＝　Channel Pruning
（※prune: 小枝を刈り取る）

どこのChannelをとるのか？

channel

channel



参照：Channel Pruning for Accelerating Very Deep Neural Network , Yihui He, ICCV

層の出力のヒストグラムの解析

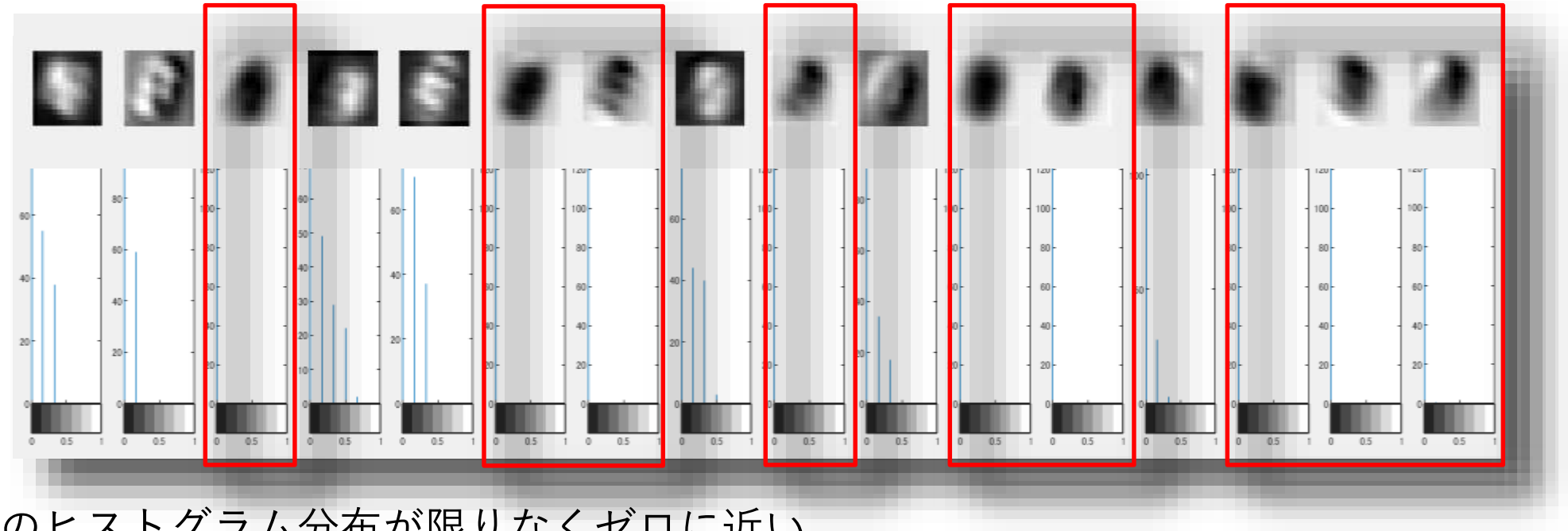
ネットワークのWeightや演算後のActivationの結果を踏まえて

再学習後のネットワーク(net)からWeightを取り出し

```
net.Layers(2).Weights
```

評価画像セットの‘conv_2‘層での出力を取り出し

```
act2 = activations(net,imdsValidation,'conv_2','OutputAs','channels');
```



□ 活性化のヒストグラム分布が限りなくゼロに近い

→ この9つのChannelの情報は判定に利用されていないのでは？

Pruningの効果(例)

AlexNet　Channel Pruning 効果

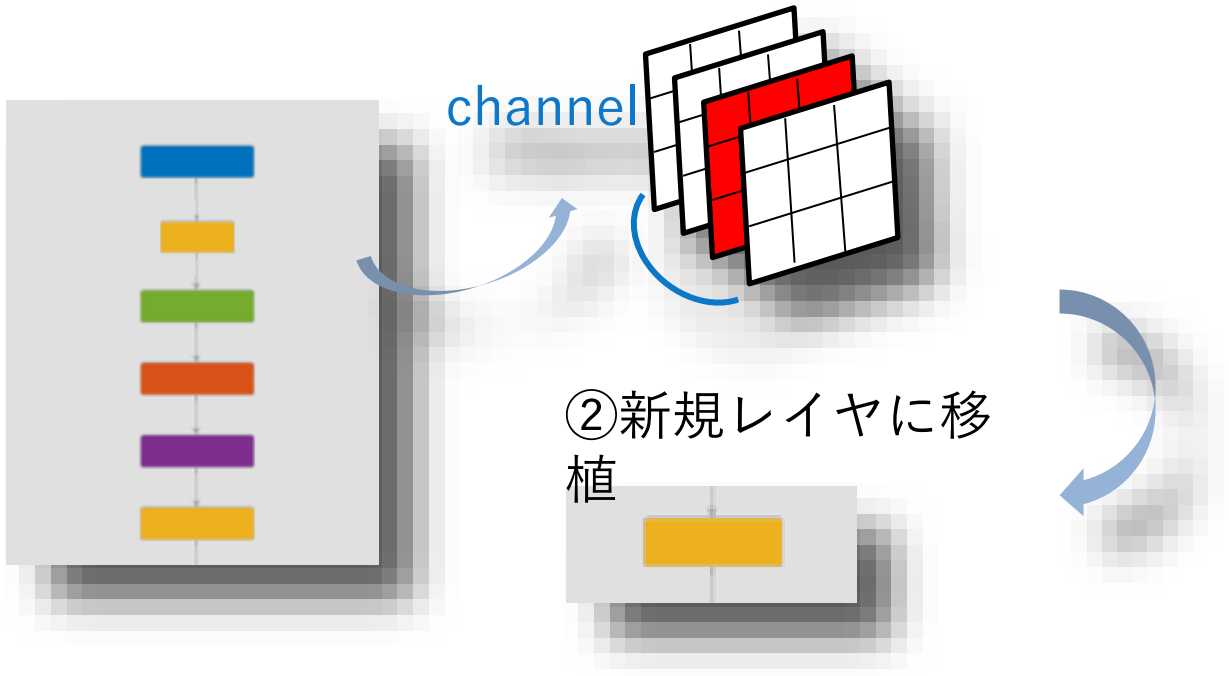
	Pruning前	Pruning後
ファイルサイズ	207MB	78MB (63%削減)
精度	93.2 %	89.4 %
推論速度(CPU)	25.1 s	19.9 s （26 %高速化）
推論速度(GPU ノートPC用)	13.0 s	10.3 s （26%高速化）
推論速度(GPU デスクトップ用)	4.32 s	4.17 s （4 %高速化）

10種類の分類。AlexNet。
推論：約700枚。Minibatchsize=128

・ファイルサイズ削減効果大
・推論速度も向上(高性能GPUを使う場合は効果少ない)
・学習時間の短縮にもつながる
・精度はトレードオフ

Pruningの流れ

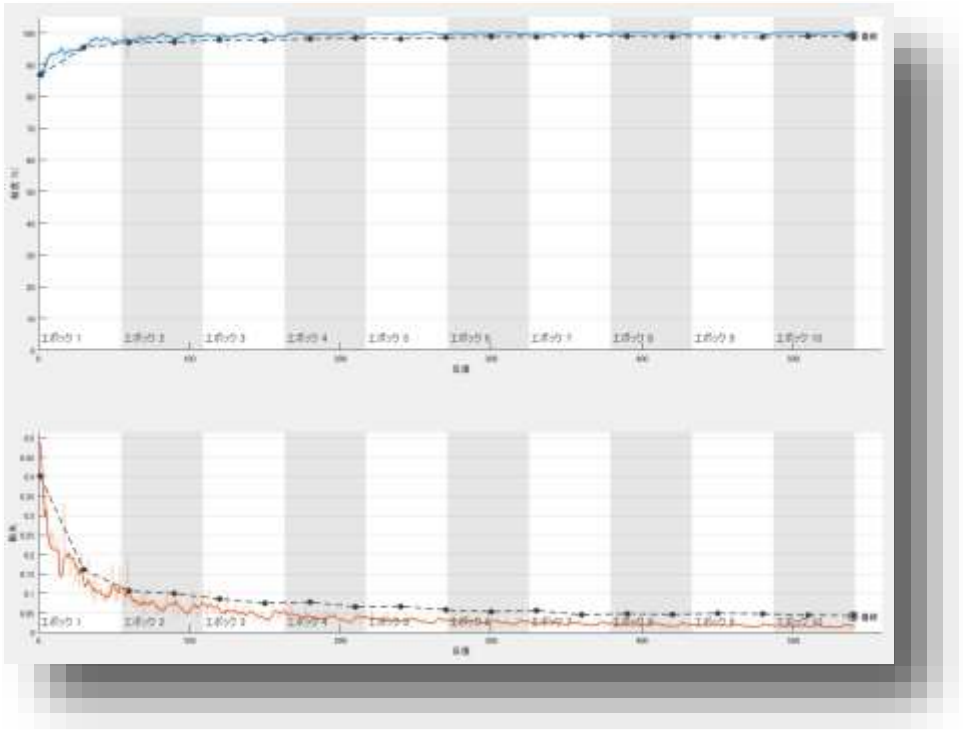
①指定ChannelのWeightの取り出し



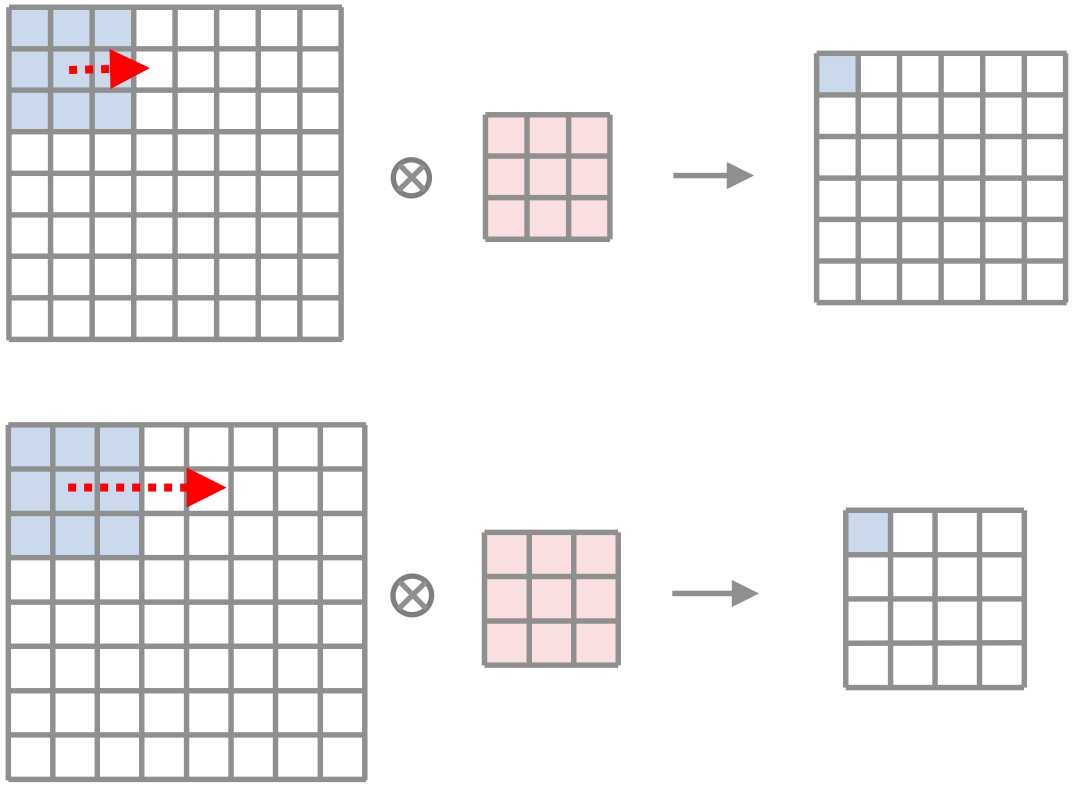
②新規レイヤに移植

③ネットワークのレイヤの取り換え

Pruning後の学習曲線



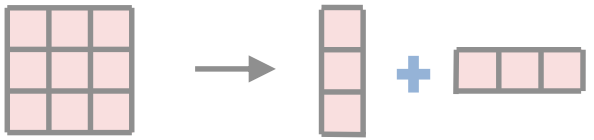
スクラッチで作成する場合
Strideを変更



- Strideを大きくすると
  - 演算量減少
  - 精度はトレードオフ

Factorization

3x3の畳み込みを
1x3 > 3x1の2層で組むことで
計算量が2/3に



アジェンダ

システム展開のためのMATLAB®活用術

システム化を踏まえたモデル選択

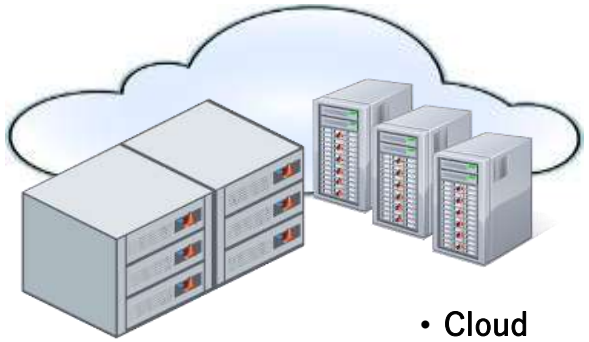
ターゲットの選択とシステム展開

システムへの統合

ネットワークの学習/評価が
終わったらシステムでの運用へ

スマホで判定

MATLAB Production Server™

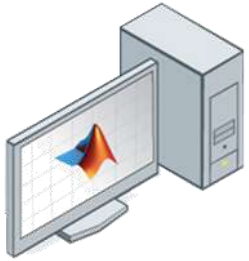


クラウドで利用

- Cloud
- Web
- Cluster
- Mobile

MATLAB Compiler™ / MATLAB Compiler SDK™

Excel Add-in
Hadoop
Standalone Application
C/C ++
Java
.NET



PCで利用

GPU Coder™

NVIDIA®
cuDNN/TensorRT™

MATLAB Coder™

MKL-DNN
(Intel® Xeon ®)

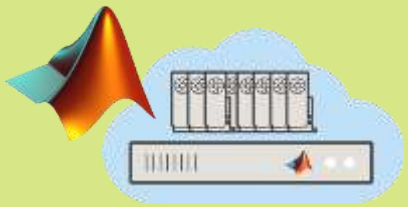
Compute Library
(Arm® Neon™)

3種類の統合へのパスをサポート
MATLABライセンスがない環境で利用可能

システムへの統合

複数のフレームワークのモデルを利用可能

Caffe MODELS
Keras
TensorFlow
ONNX

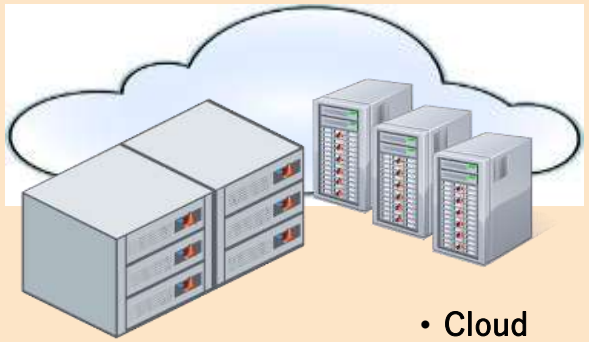


アプリケーション

複数のターゲット

スマホで判定

MATLAB Production Server™

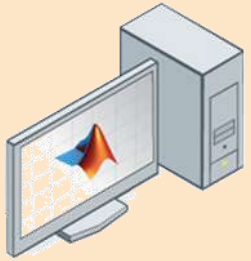


クラウドで利用

- Cloud
- Web
- Cluster
- Mobile

MATLAB Compiler™ / MATLAB Compiler SDK™

Excel Add-in
Hadoop
Standalone Application
C/C ++
Java
.NET



PCで利用

GPU Coder™

NVIDIA®
cuDNN/TensorRT™

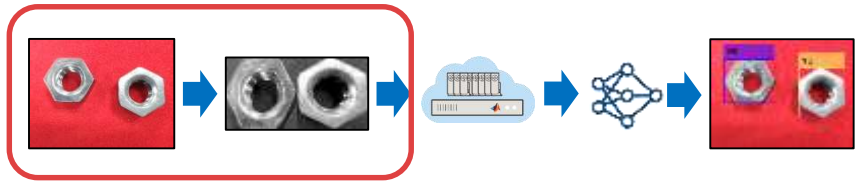
MATLAB Coder™

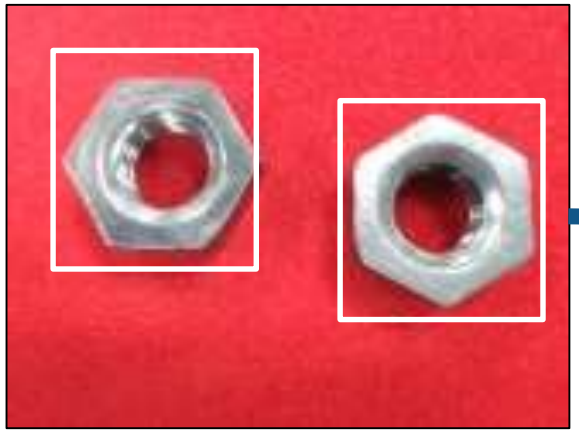
MKL-DNN
(Intel® Xeon ®)

Compute Library
(Arm® Neon™)

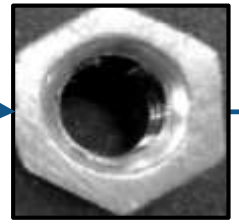
アプリケーション化のための機能

前処理

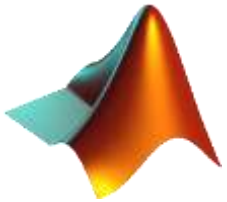




候補領域抽出



リサイズ



Caffe MODELS

Keras

TensorFlow

ONNX

特徴抽出 + 分類

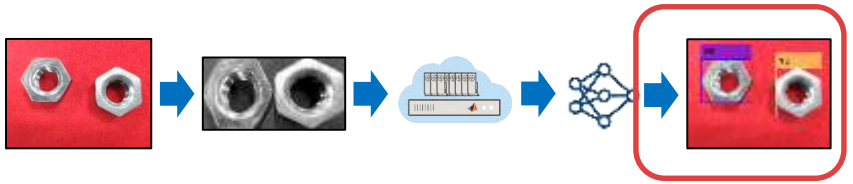
CNN

候補領域ごとにCNNで分類

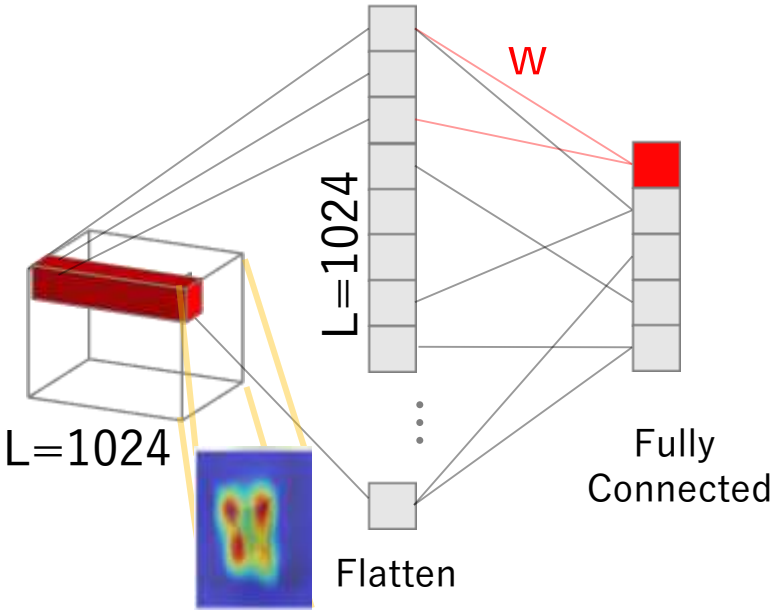
前処理を効率的に統合してアルゴリズム構築

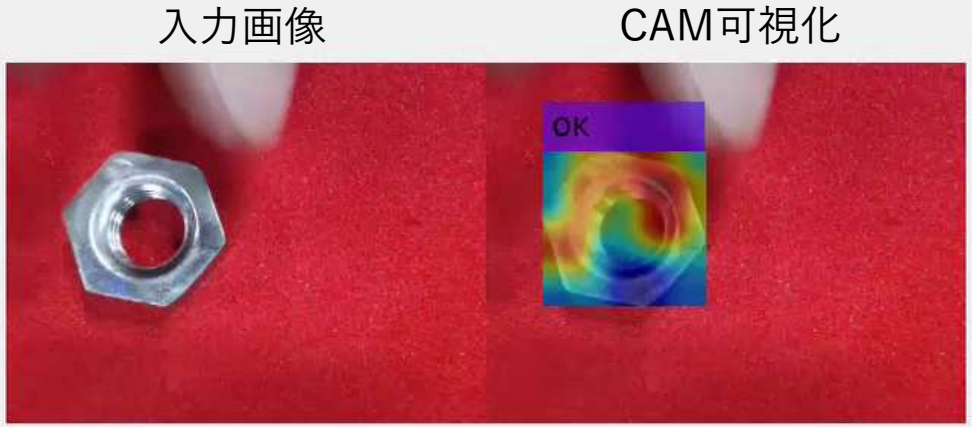
判定結果の可視化

後処理



CAM(Class Activation Mapping)による特徴可視化





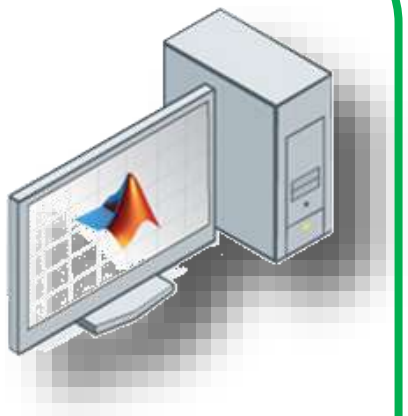

```
dotProduct =
bsxfun(@times,imageActivations,weightVector);
classActivationMap = sum(dotProduct,3);
```

OK→ ナット表面全体に注目
NG→ キズの部分に反応

学習済みネットワークが
どのような特徴をとらえているかを可視化
後処理も含めてCUDAコード生成可能

参考資料:
Learning Deep Features for Discriminative Localization
Bolei Zhou etc, Computer Science and Artificial Intelligence Laboratory, MIT

MATLAB開発における実行システムの選択



PC & GPU

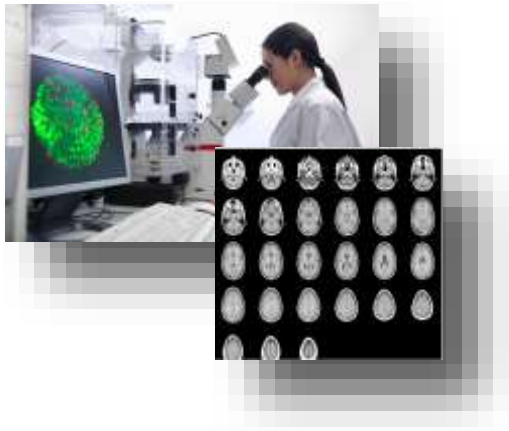
サーバー/クラウド

組み込みGPU

組み込みCPU

PCで使う（MATLAB Runtimeを利用）

MATLAB Compiler™ / MATLAB Compiler SDK™

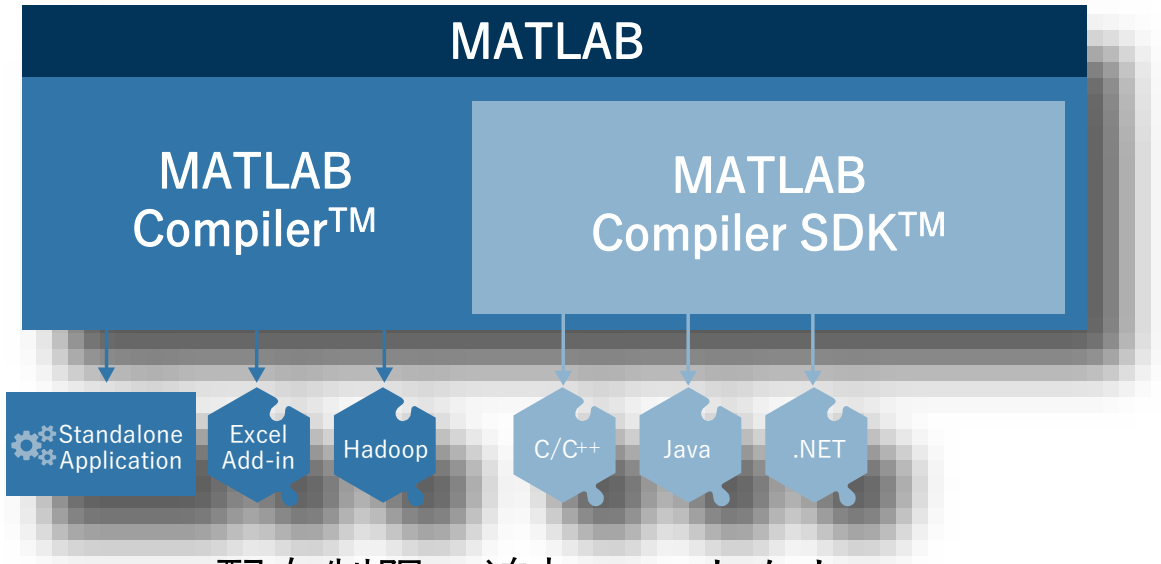


メリット

・OS選択（Windows/Mac/Linux）
・GUI操作
・MATLABで開発したものをそのまま実行可能
・実行ファイル・ライブラリとして配布
・GPU演算 (Parallel Computing Toolbox)

利用場面

・診断 (GUI活用)
・既存システムに統合(.NET/C他)
・マシンコストをかけられる (台数少)
・モデルの変更が多い(プロトタイプ)



・配布制限・追加コストなし
・無償のRuntime

プロトタイプ段階では現場専門家とAI開発者の協力が必要

判断の経験をもつ現場専門家

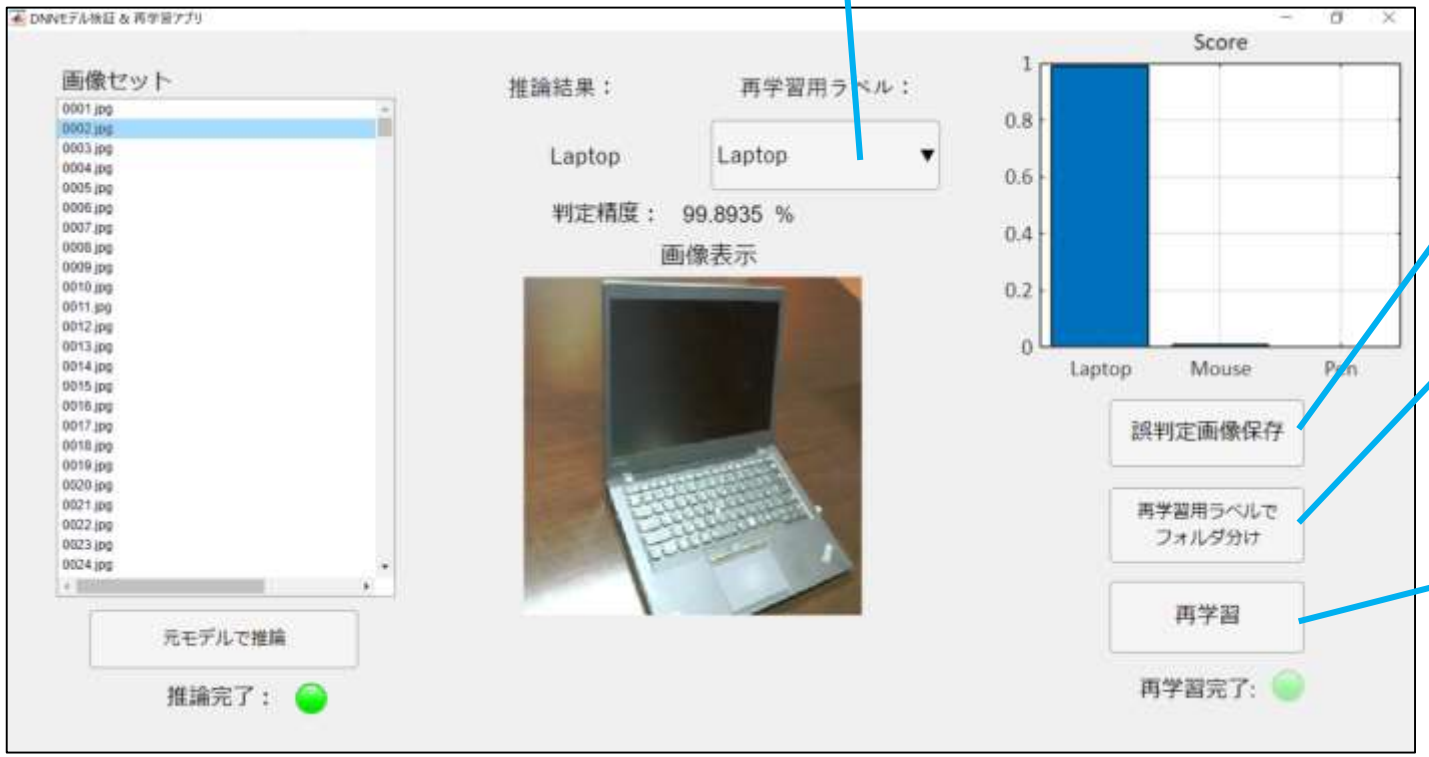
「Deep Learningの
詳しいことはわからない」

AI 開発者

「これはどっちに分類？現場の
職人さんじゃないと。。。」

必要な操作をGUIにしEXE/ライブラリ/WebApp配布

現場のスペシャリスト
が推論間違いを修正



誤判定画像
のとりだし

修正後のラベル
でフォルダ分け

再学習

推論精度の確認、ラベルの訂正、再学習までをマウス操作で

Standalone
Application

PCで使う（Intel MKL-DNN）

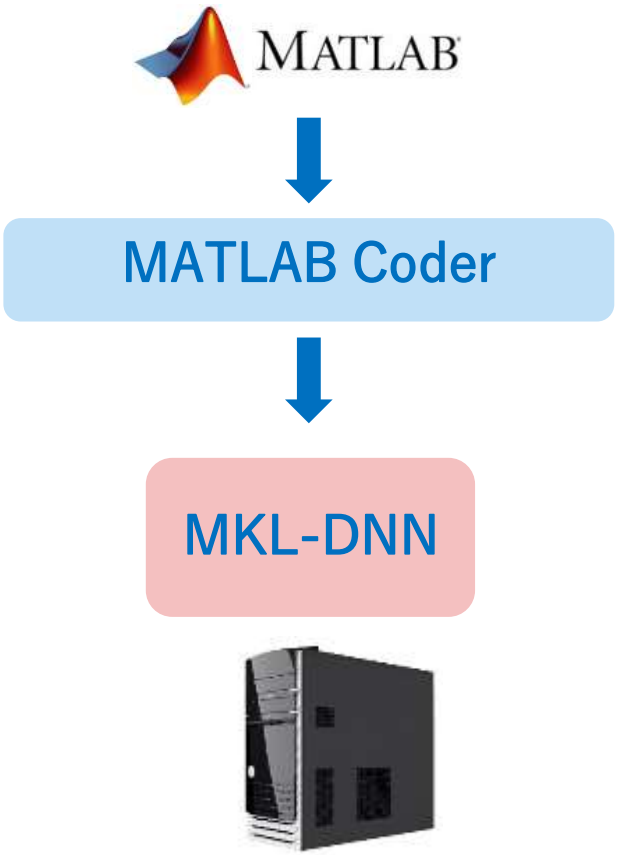
MATLAB Coder™

メリット

・OS選択（Windows/Linux）
・GPUがない環境下で高速化
・Runtime不要

利用場面

・比較的規模の大きい検査装置
（ライン検査等）
・既存システムに統合(.NET/C他)



Intel® Core™ i7 + MKL-DNNで実行した様子

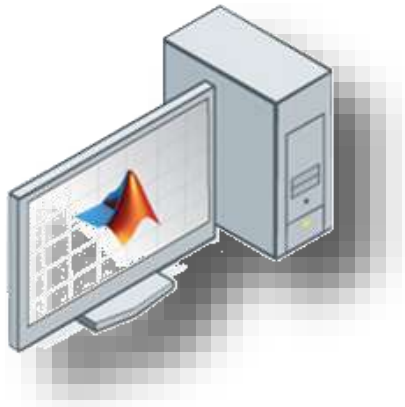
CVST Codegen + MKL-DNN AlexNet

CVST Codegen + MKL-DNN YOLO

OpenCV + MKL-DNN AlexNet

OpenCV + MKL-DNN ResNet-50

MATLAB開発における実行システムの選択



PC & GPU

サーバー/クラウド

組み込みGPU

組み込みCPU

サーバー/クラウドで使う

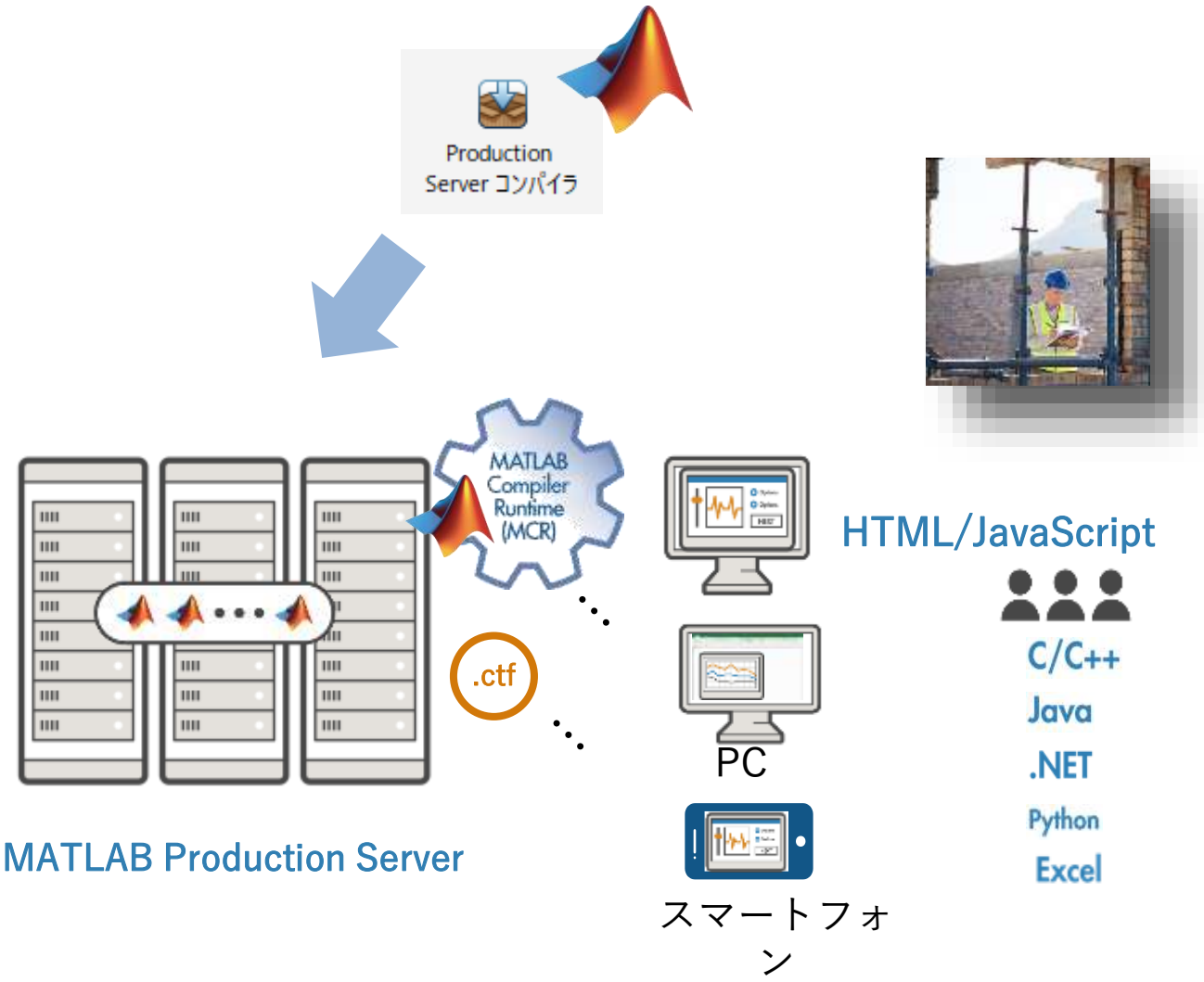
MATLAB Production Server™

メリット

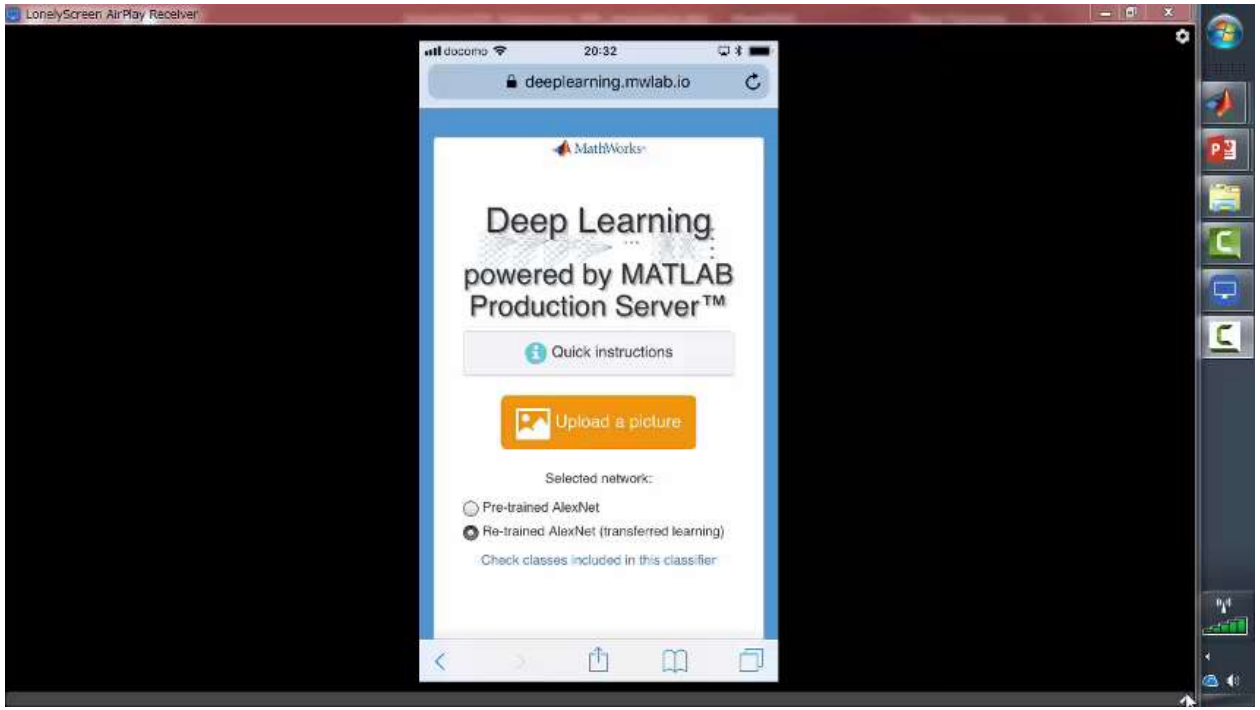
・RestfulなServerを立てられる
（WEB ServerからhttpリクエストⅠ）
・クライアント側にRuntimeが不要
・同時アクセス/複数ワーカーに対応

利用場面

・クラウドサービスのバックエンド
・スマートフォンからのアクセス
・モデルの集中管理&実行



Demo：　スマホでディープラーニング



1. QRコードでページにアクセス
2. ネットワークを選択して写真をとる
3. 数秒後分類結果の表示

https://deeplearning.mwlab.io

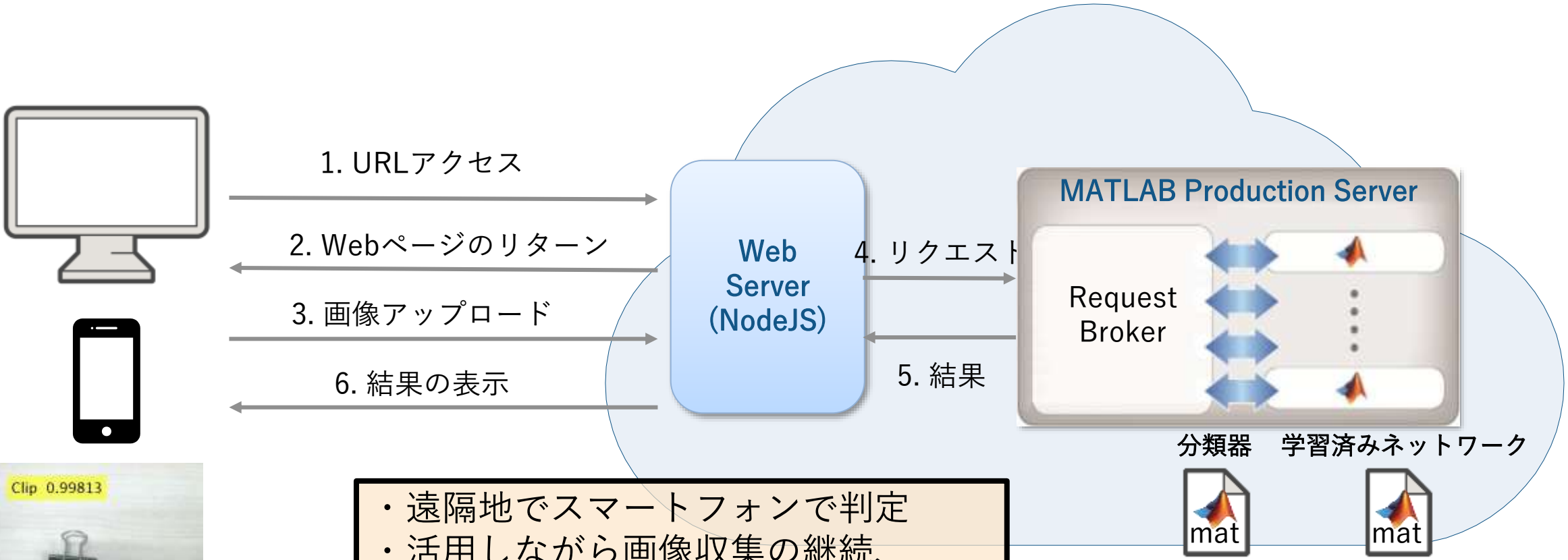
●ネットワークは２種類から選択

1. Pre-trained AlexNet
   -1000種類
2. Re-trained SqueezeNet

□ボトル	□顔
□カン	□ノートPC
□コイン	□マウス
□お皿	□ペン
□コップ	□腕時計

クラウドへの展開

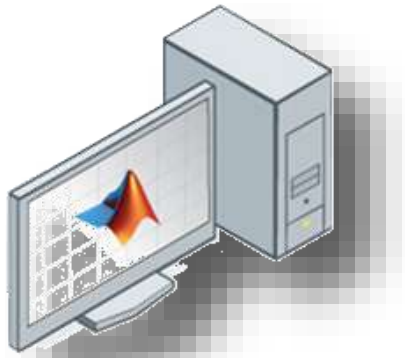
MATLAB® Production Server™



・遠隔地でスマートフォンで判定
・活用しながら画像収集の継続、
　ネットワークの再学習⇒精度向上

短期間でシステムとして使い始めることが重要です！！

MATLAB開発における実行システムの選択



PC & GPU

サーバー/クラウド

組み込みGPU

組み込みCPU

組み込みGPU

GPU Coder™

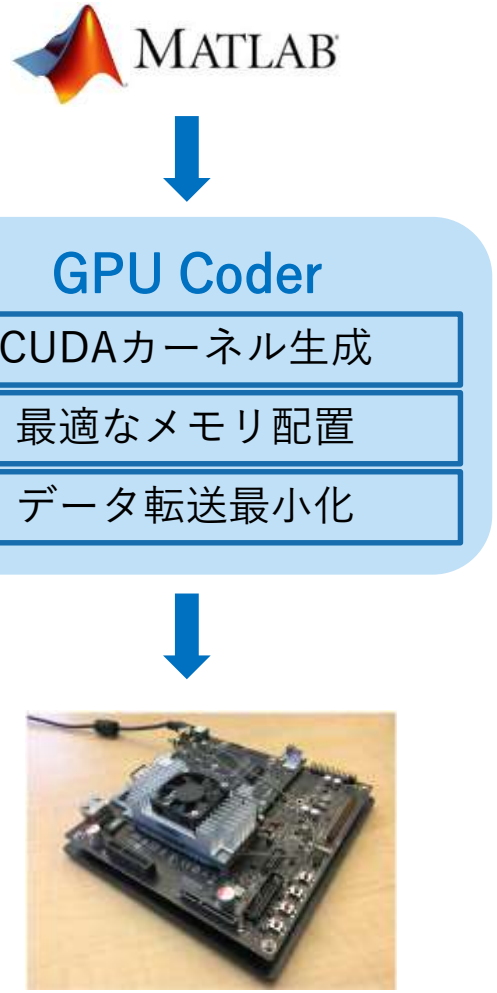
メリット

・高速並列演算に特化
(ディープラーニングで費用対効果〇)
・I/O

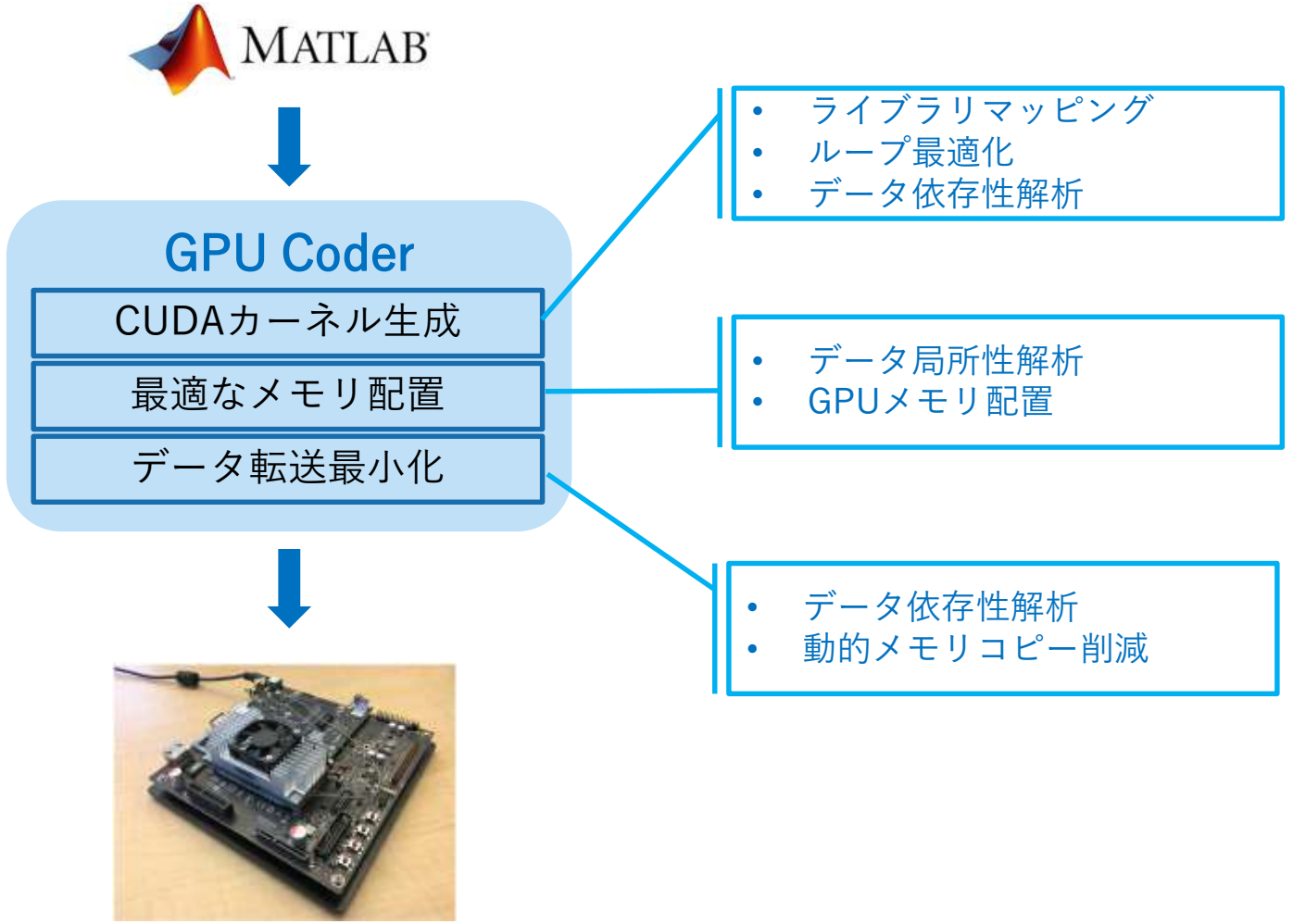
利用場面

- 高速なエッジ処理
  - ライン検査
- 自律型組み込みシステム向け
  - 自動運転
  - 自律型ロボット
  - ドローン

関数解析機能により効率の良いコード生成



GPU Coder : 高度な関数解析機能が効率の良いコード生成を実現



MATLABからCUDA C/C++コードを自動生成します

CPU-GPU間のデータ転送最適化について

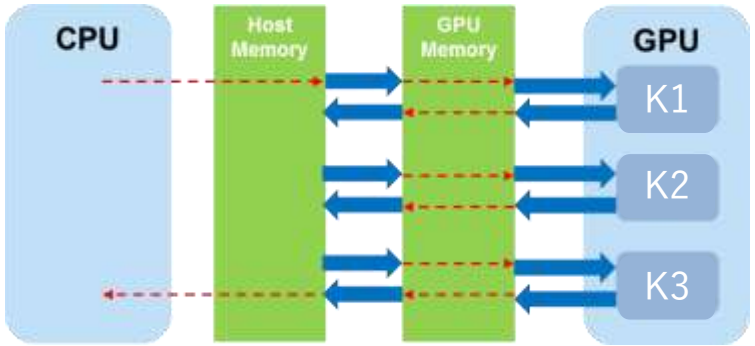
```
A = …
…
for i = 1:N
   … A(i)
end
…

imfilter
…
```

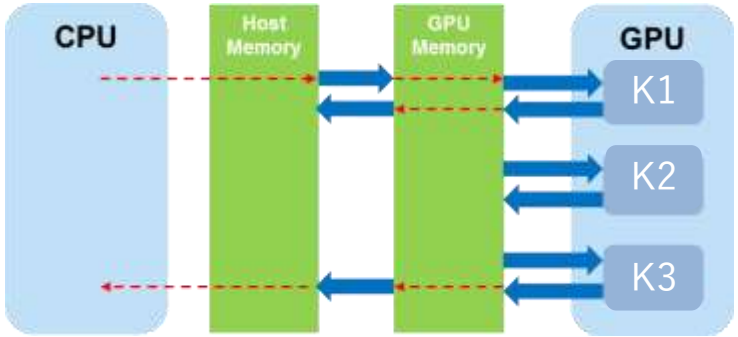
```
A = …
…
cudaMemcpyHtoD(gA, a);
kernel1<<<…>>>(gA)
cudaMemcpyDtoH(…)
…
cudaMemcpyHtoD(…)
imfilter_kernel(…)
cudaMemcpyDtoH(…)
…
```

理想的なmemcpyの配置位置は?

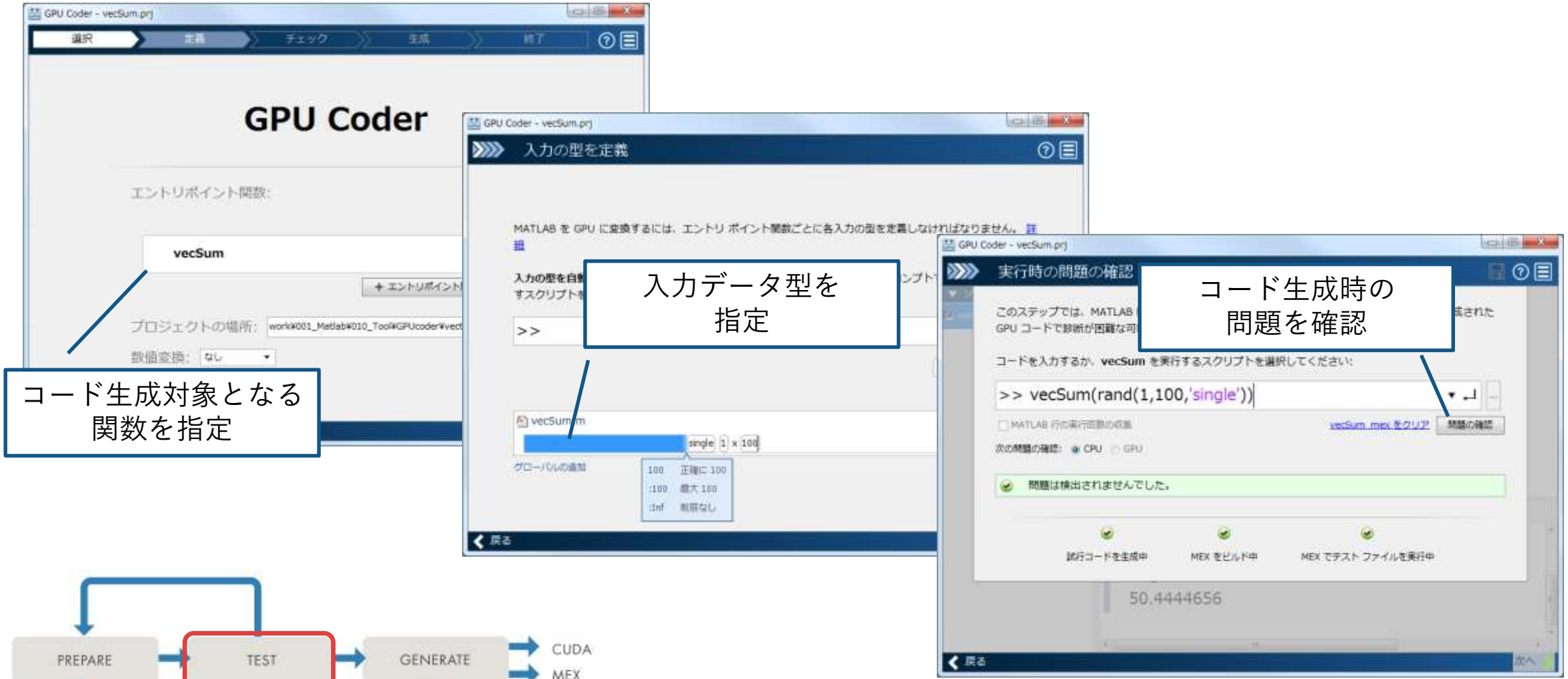
都度転送



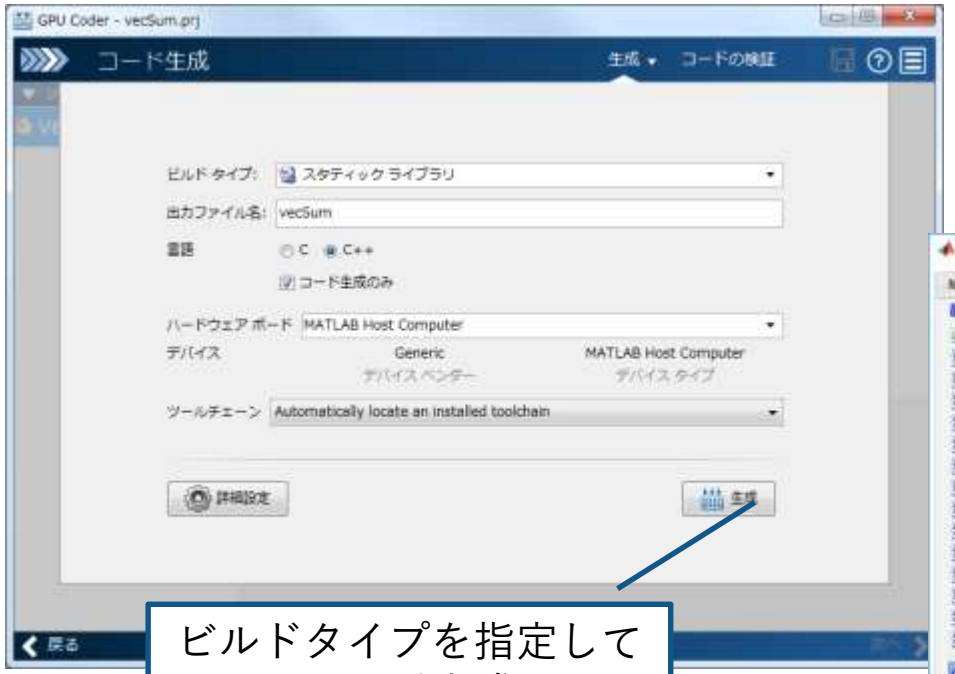
最適なデータ転送



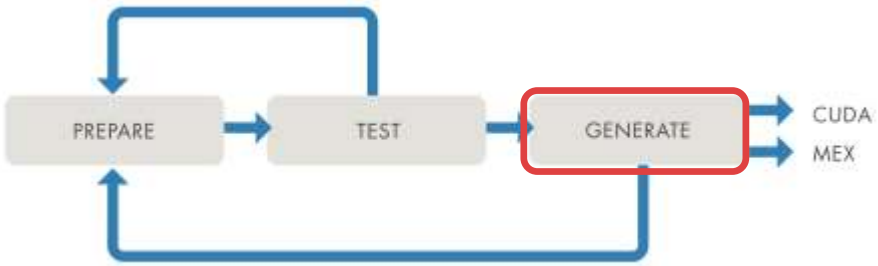
GPU Coder : 専用Appでコード生成可能か確認



GPU Coder : コード生成とレポート確認



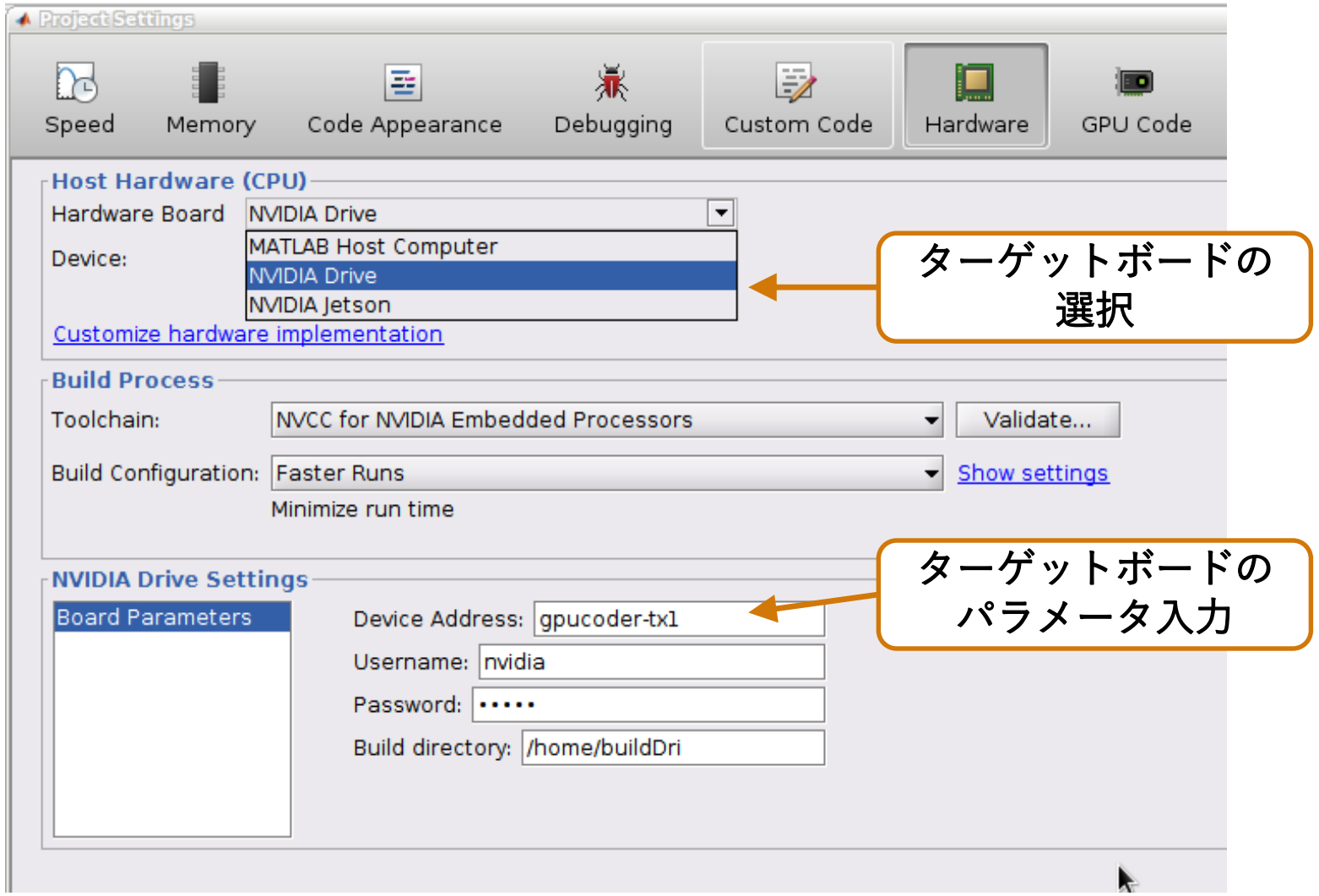
ビルドタイプを指定して
コード生成



組み込みGPUでのアルゴリズム調整は？

NVIDIA Hardware Support Package (HSP)

R2018b



ターゲットボードの選択

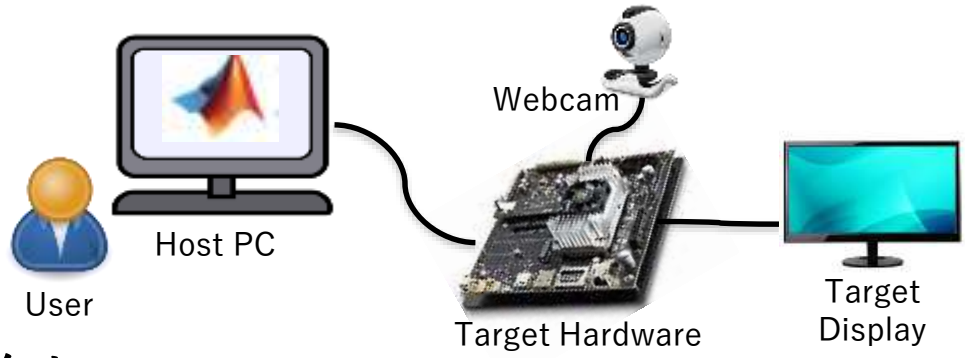
ターゲットボードのパラメータ入力

MATLABから直接ボードにアクセス可能

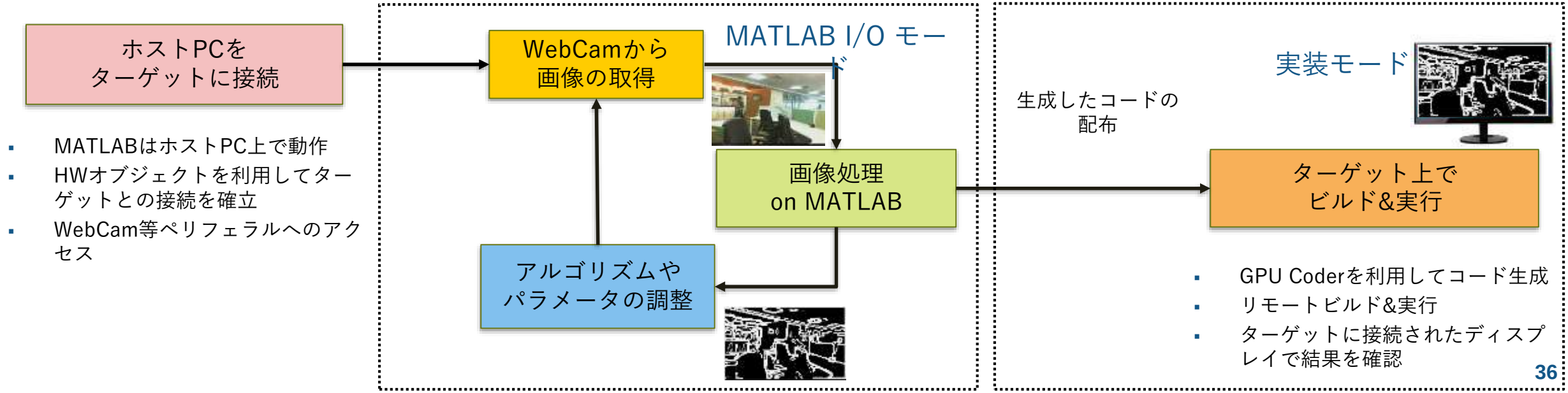
Jetson / Drive platform

NVIDIA hardware support package

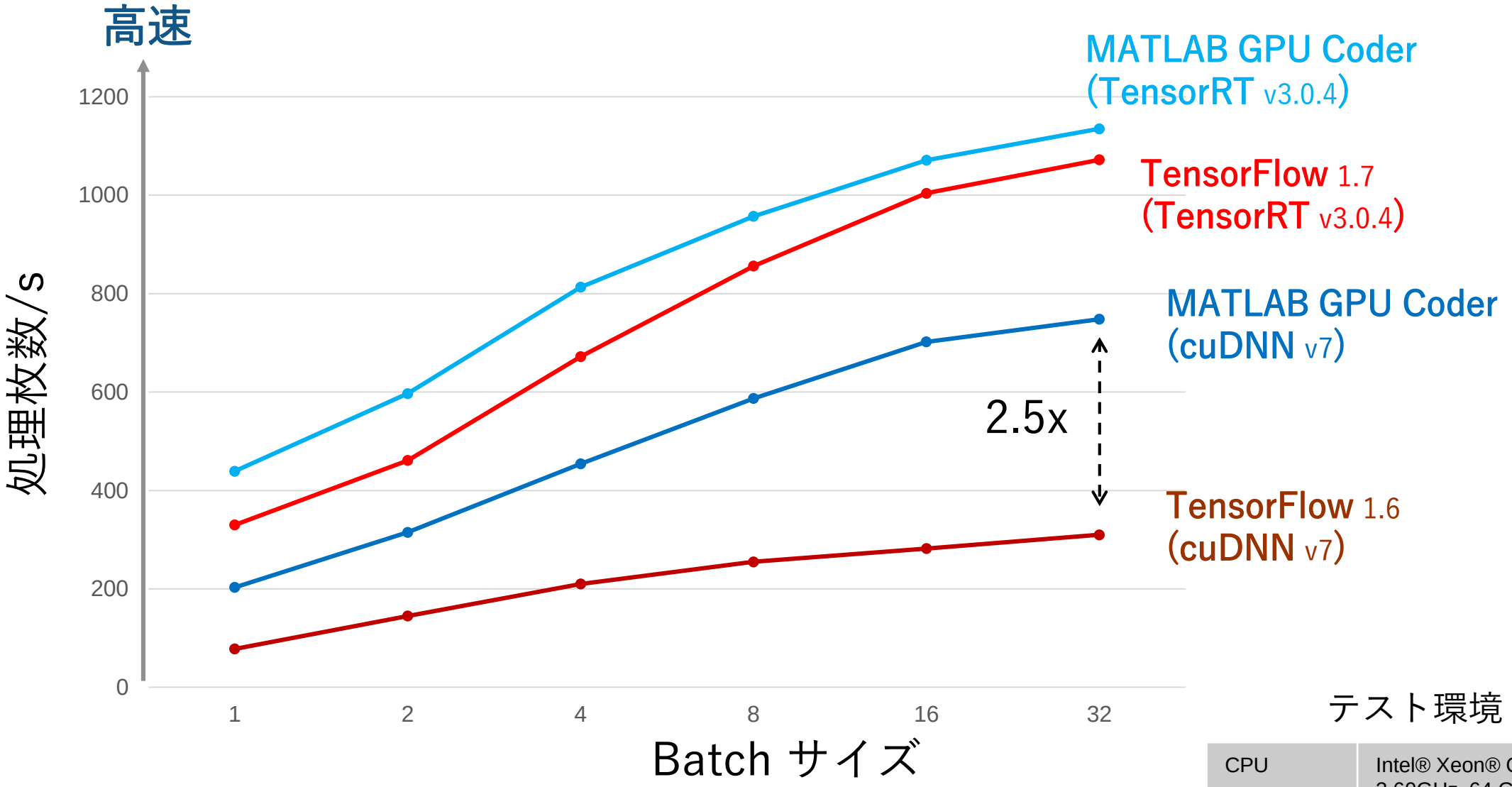
R2018b



- NVIDIA GPUハードウェアへの
  アクセスと実装がより容易に行えます。
  - **ターゲットハードウェアに接続されたセンサへのアクセス**
  - **GPU Coderにより生成されたファイル類のコピー、リモートビルド**



Resnet-50の推論速度

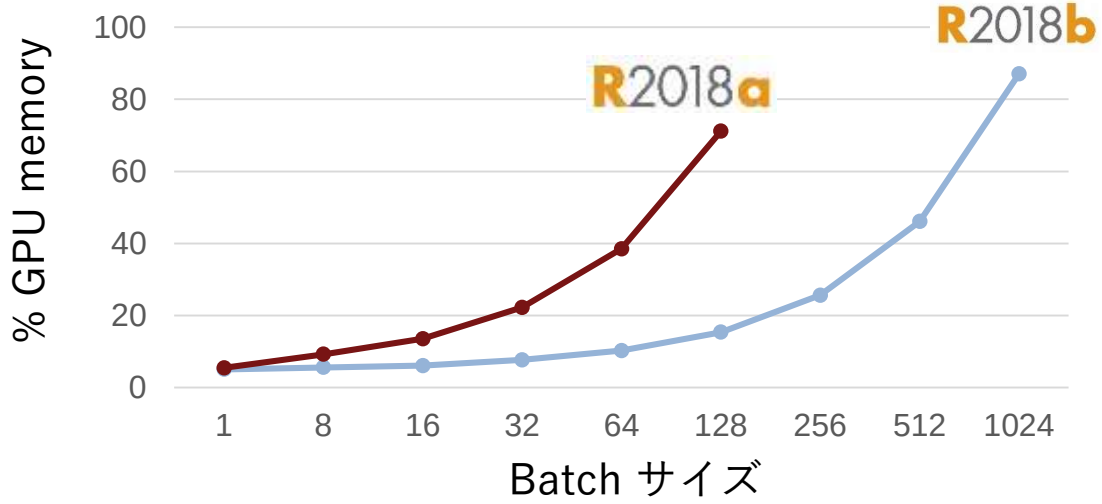


テスト環境

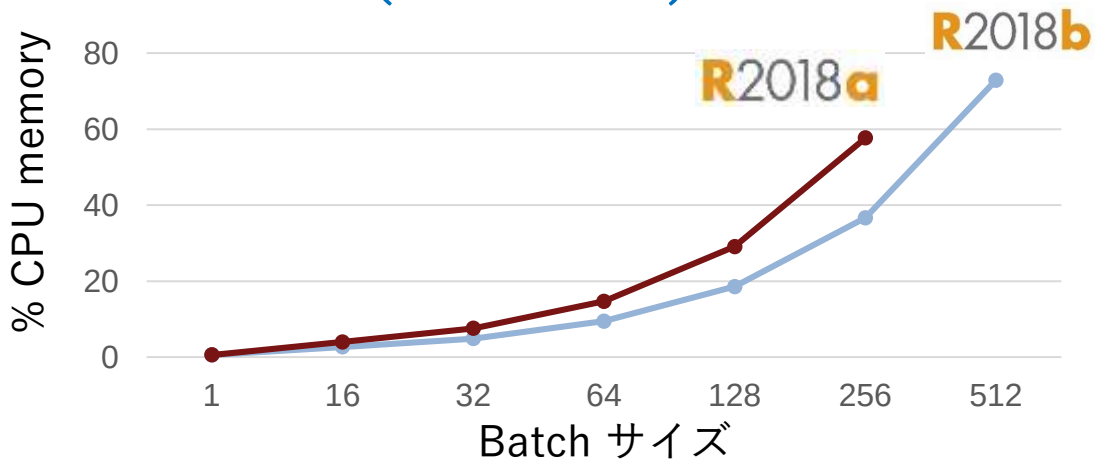
CPU	Intel® Xeon® CPU E5-1650 v4 @ 3.60GHz, 64 GB RAM
GPU	Pascal Titan Xp, 12 GB RAM

R2018b GPU Coderにおける最適化
バッファーの最小化

Resnet-50 （ TitanXP ）

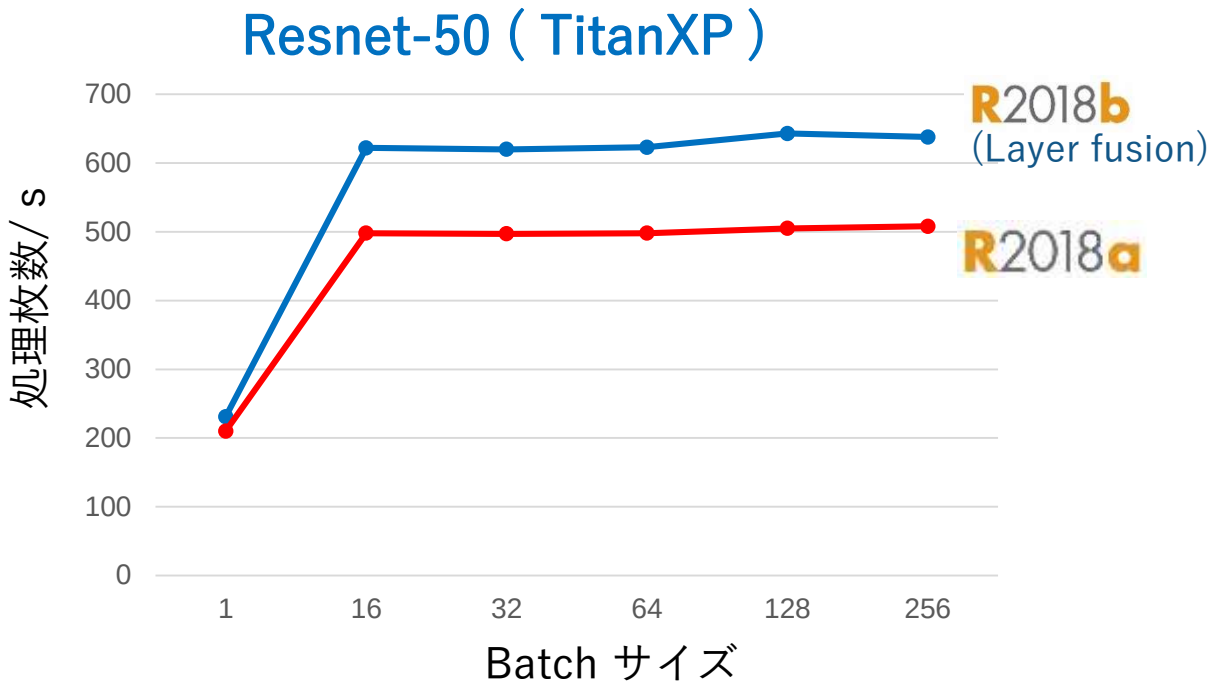


Resnet-50 （ Intel Xeon ）



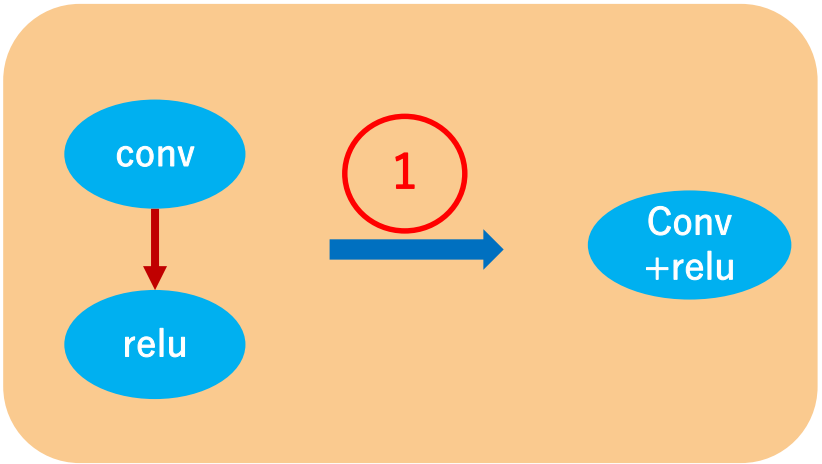
バッファーの最小化

R2018b GPU Coderにおける最適化: レイヤーフュージョン

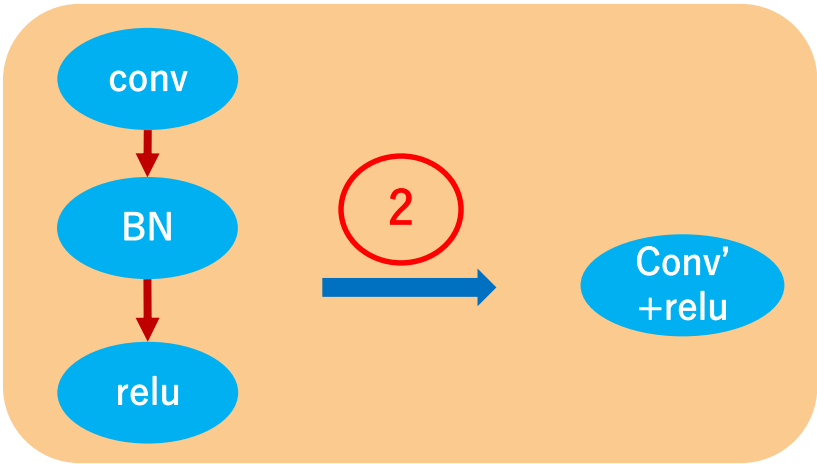


バッファーの最小化

レイヤーフュージョン

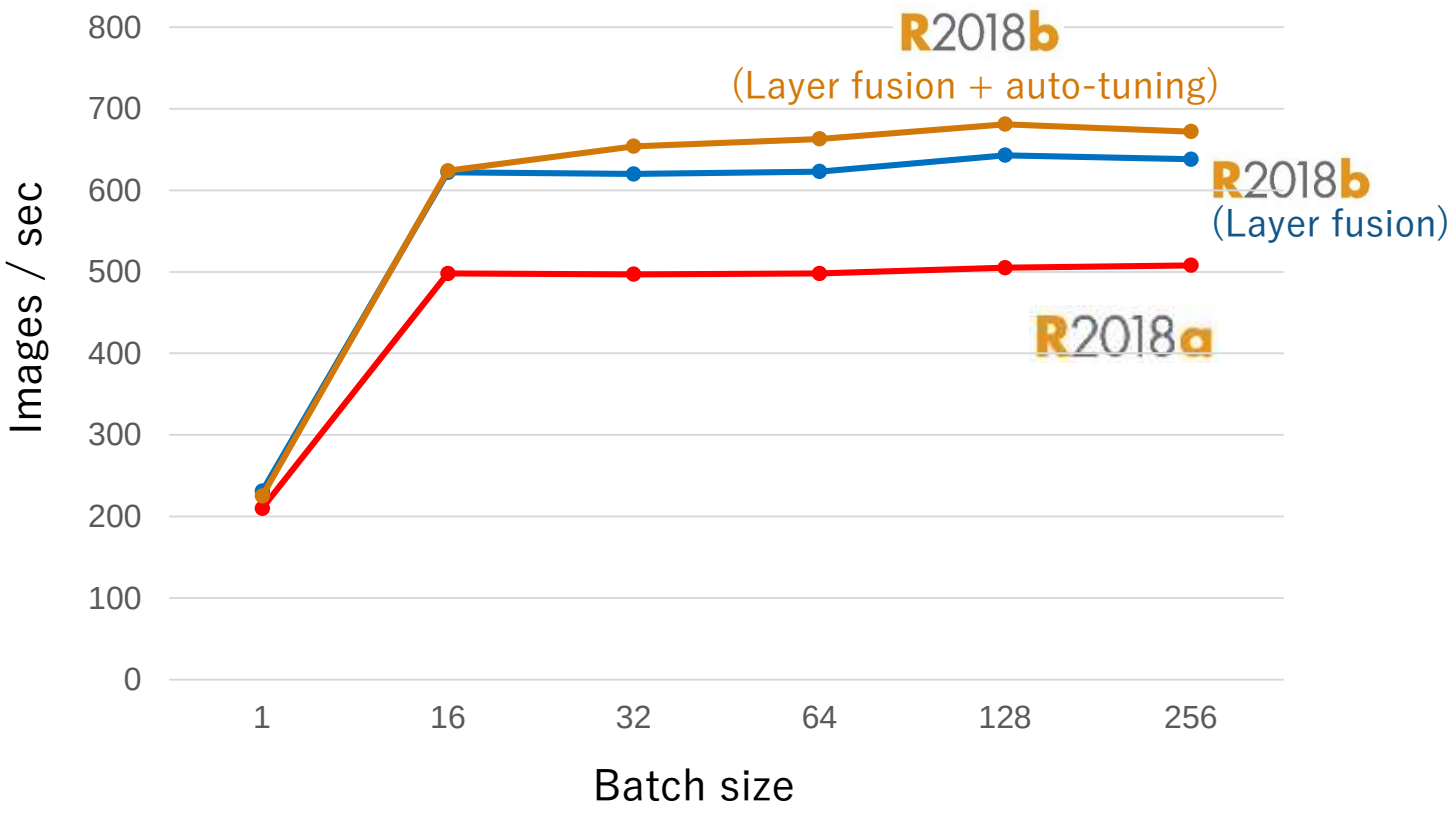


複数レイヤを
一つに置き換え



R2018b GPU Coderにおける最適化: 自動調整

Resnet-50 (TitanXP)

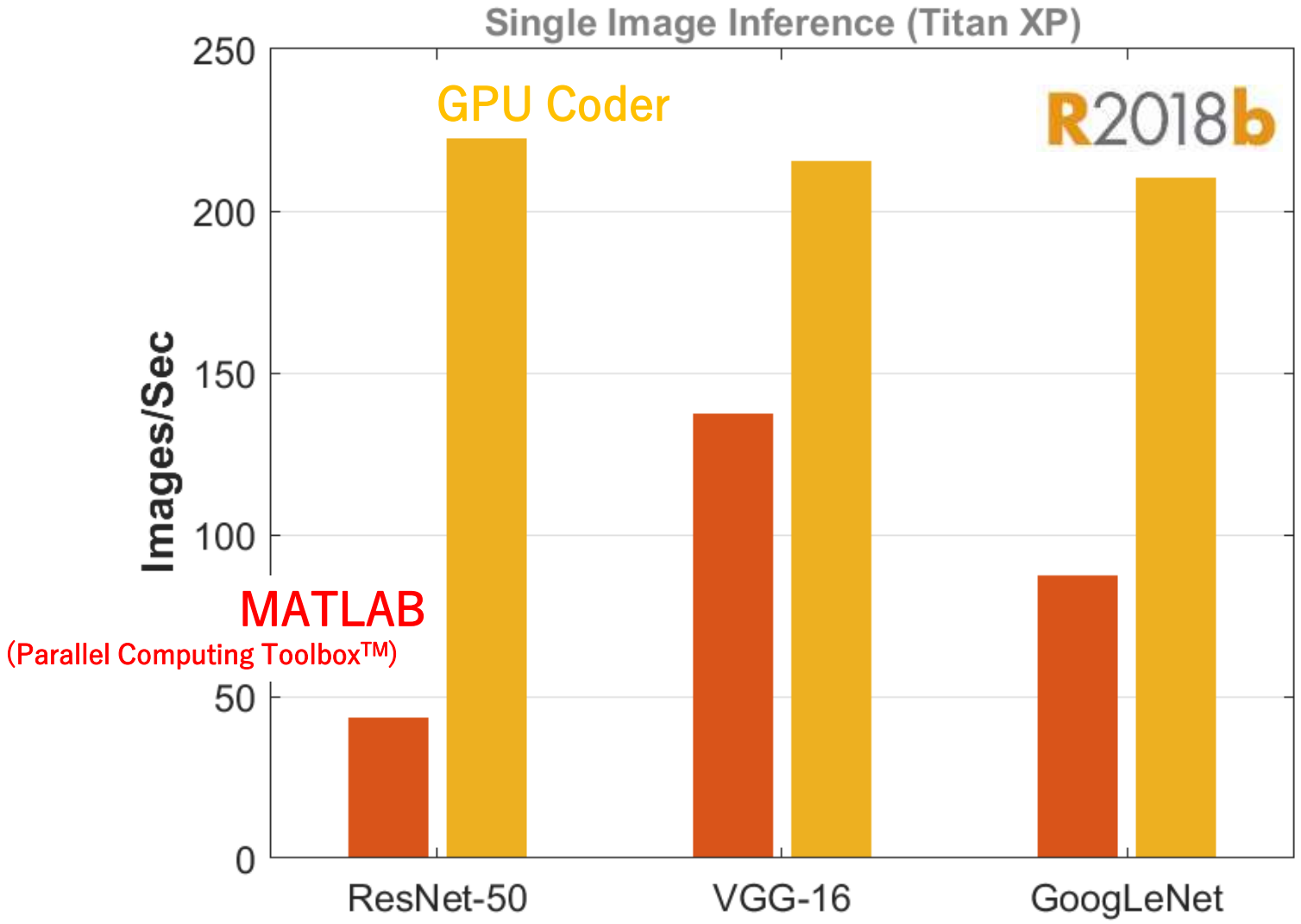


バッファーの最小化

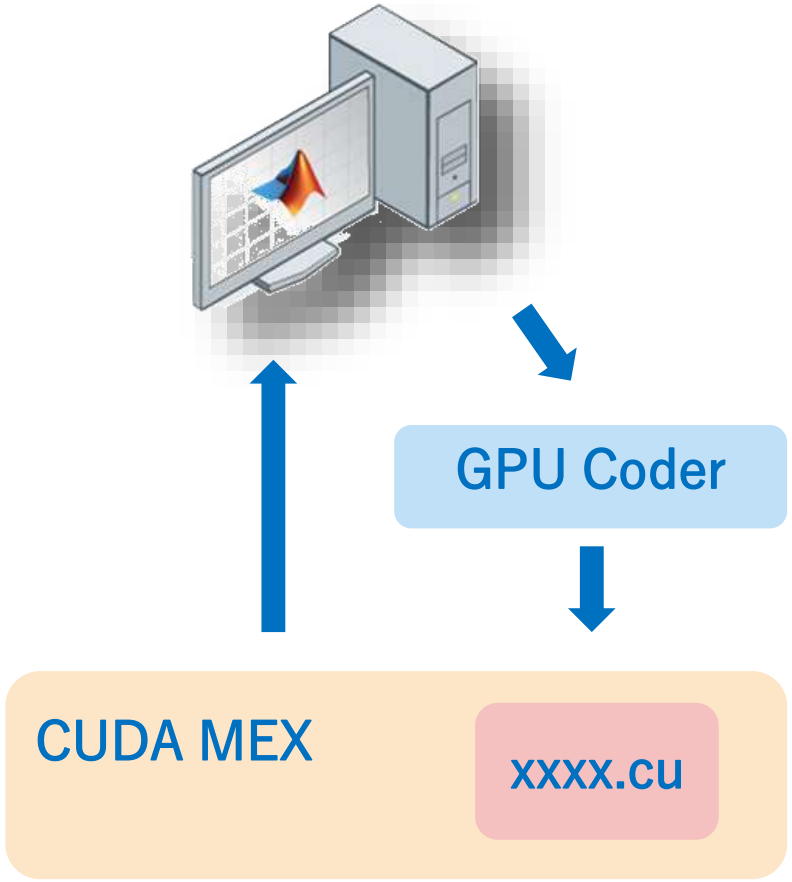
レイヤーフュージョン

自動調整

GPU CoderはPCの推論も高速化

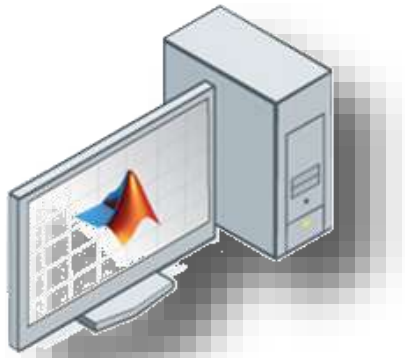


Intel® Xeon® CPU 3.6 GHz – NVIDIA® libraries: CUDA®9 - cuDNN 7



生成したCUDAコードをMEX化してMATLABに取り込み

MATLAB開発における実行システムの選択



PC & GPU

サーバー/クラウド

組み込みGPU

組み込みCPU

CPU用ライブラリの活用 (Arm® Compute library)

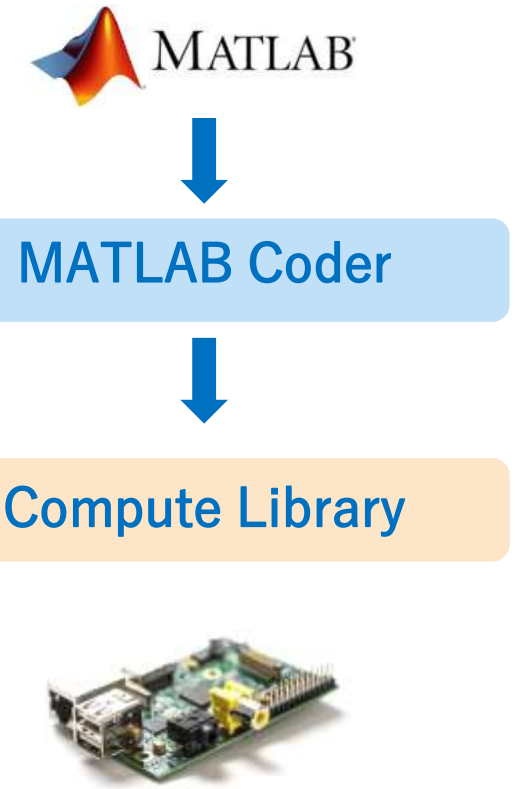
MATLAB Coder™

メリット

・低コストデバイス
・低消費電力
・サイズ
・I/O

利用場面

- 多数のエッジ処理が必要
- 高速性が不要
  - 外観検査(高速性不要)
  - IoT
    - メンテナンス
    - 環境/気象



ターゲットハードウェアとコード生成方法の対応について

Deep Learning Networks

codegen

もしくは、
GPU Coder App

cncodegen

・CNN+画像前処理、後処理も含めて
コード生成可能

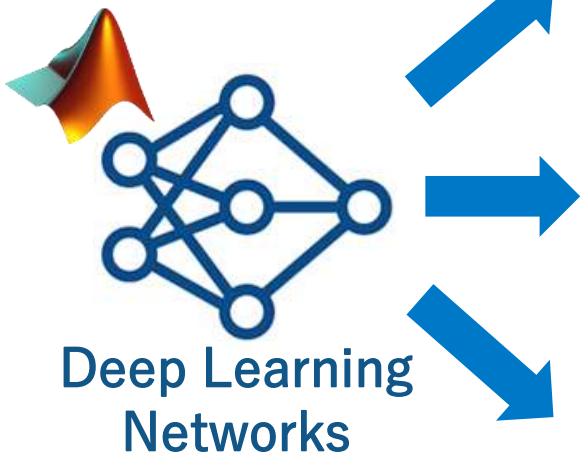
Intel
MKL-DNN
Library

NVIDIA
TensorRT &
cuDNN
Library

Arm
Compute
Library

・CNNの部分のみコード生成可能
・ターゲットハードウェアを複数選択可能

パラメータ変更で簡単なターゲット変更



Intel® MKL-DNN Library

```
cnncodegen(net,'targetlib','mkl-dnn');
```

NVIDIA® TensorRT™ & cuDNN Library

```
cnncodegen(net,'targetlib','tensorrt');
```

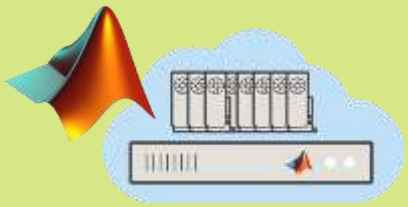
Arm® Compute Library

```
cnncodegen(net,'targetlib','arm-compute');
```

システムへの統合

複数のフレームワークのモデルを利用可能

Caffe MODELS
Keras
TensorFlow
ONNX

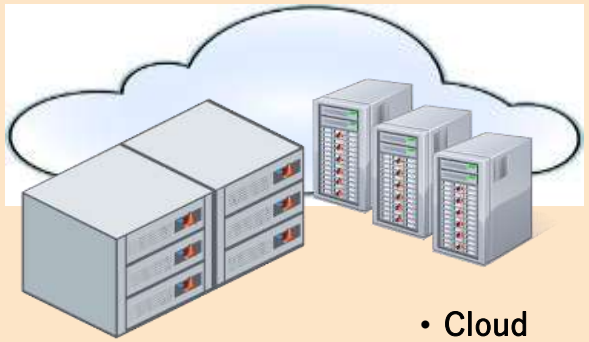


アプリケーション

複数のターゲット

スマホで判定

MATLAB Production Server™

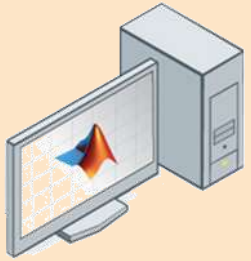


クラウドで利用

- Cloud
- Web
- Cluster
- Mobile

MATLAB Compiler™ / MATLAB Compiler SDK™

Excel Add-in
Hadoop
Standalone Application
C/C ++
Java
.NET



PCで利用

GPU Coder™

NVIDIA®
cuDNN/TensorRT™

MATLAB Coder™

MKL-DNN
(Intel® Xeon ®)

Compute Library
(Arm® Neon™)

まとめ

ディープラーニングシステム展開におけるMATLAB活用のメリット

PC、エッジデバイス、クラウドなど複数のターゲットパスに対応

前処理、可視化を含めて開発し統合が可能

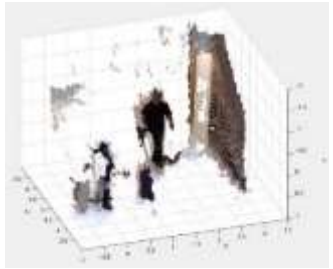
推論高速化にはGPU Coderが効果的 (TensorRTに対応)

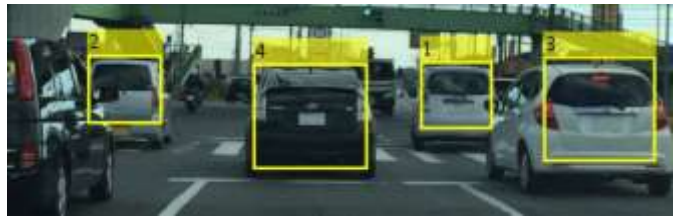
プロジェクトごとにターゲットは変わります。
柔軟な選択が可能なMATLABをご活用ください。

Next Steps : 画像処理・ディープラーニング無料セミナー

申し込みは弊社ウェブサイトより
https://jp.mathworks.com/company/events/seminars/ipcv-tokyo-2607088.html

- 日時：2018年11月20日 13:30-17:00
- 場所：秋葉原UDX Next-1(JR秋葉原駅より徒歩2分)
  （アクセス：http://www.udx-n.jp/access.html
- **画像処理、ディープラーニング関連機能をご紹介！**
  - MATLABではじめる画像処理ワークフロー
  - 例題で実感するMATLABの画像処理・ディープラーニング
  - 自動運転・ロボティクス領域における画像・LiDAR点群処理と実装

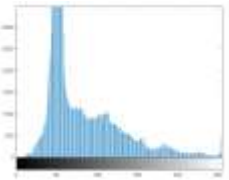




GPU Coder関連製品：画像処理･コンピュータービジョン

Image Processing Toolbox™

- コーナー、円検出
- 幾何学的変換
- 各種画像フィルタ処理
- レジストレーション（位置合せ）
- セグメンテーション（領域分割）
- 画像の領域の定量評価



Computer Vision System Toolbox™

- カメラキャリブレーション
- 特徴点･特徴量抽出
- 機械学習による物体認識
- 動画ストリーミング処理
- トラッキング
- ステレオビジョン･３Ｄ表示

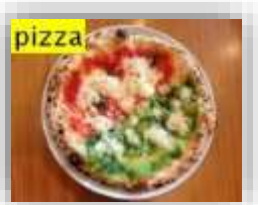
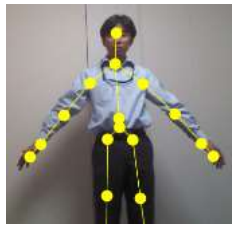



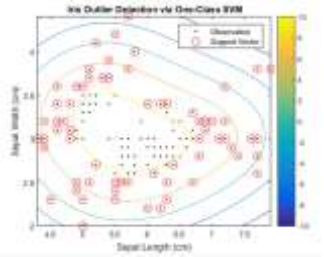
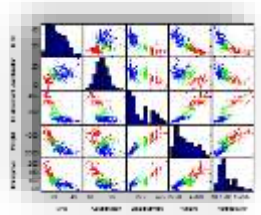
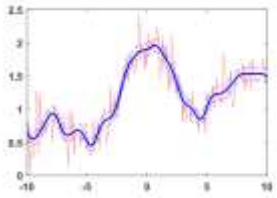
Image Acquisition Toolbox™

- デバイスから画像、動画直接取り込み
  - フレームグラバボード
  - DCAM, Camera Link®
  - GigE Vision®, Webカメラ
  - Microsoft® Kinect® for Windows®


Statistics and Machine Learning Toolbox™

- 機械学習
- 多変量統計
- 確率分布
- 回帰と分散分析
- 実験計画
- 統計的工程管理

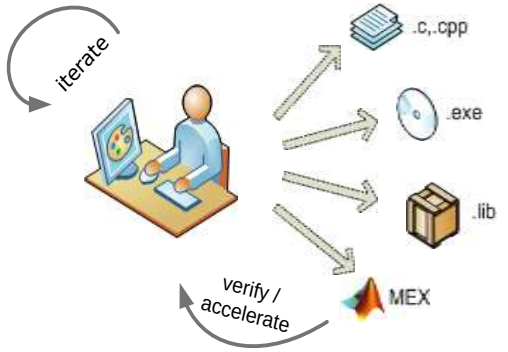




GPU Coder関連製品

GPU Coderに
必須となります

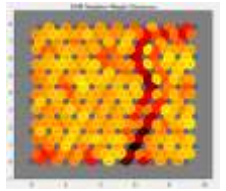
MATLAB Coder™

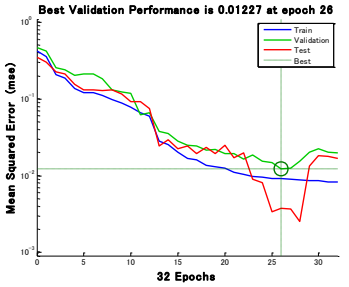
- MATLABプログラムからC/C++コードを生成
- MATLAB上で、アルゴリズム開発から実装までフローを統合

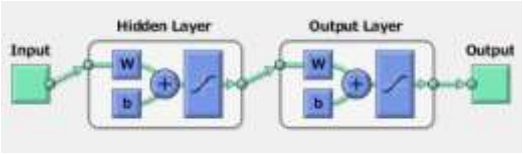


Neural Network Toolbox

- ニューラルネットワークの構築、学習
- データフィッティング
- クラスタリング
- パターン認識
- 深層学習
- GPUによる計算の高速化







Parallel Computing Toolbox

- MATLAB & Simulink と連携した並列処理
- 対話的な並列計算実行
- GPGPU による高速演算
- ジョブおよびタスクの制御

Embedded Coder®

- MATLABプログラム/Simulinkモデルから組込み用C/C++コードを自動生成