CPUだけで行けちゃうんです！
OPENVINO™で実現するAI推論の高速化

インテル株式会社
AI テクニカルソリューションスペシャリスト

大内山 浩

2019 年 10月9日

xPython Meet Up & Conference 2019

法律的な免責条項

本資料には、製品、サービス、開発プロセスに関する情報が記載されています。 ここに記載されているすべての情報は、予告なく変更されることがあります。最新の予測、スケジュール、仕様、およびロードマップをご希望の方は、インテルの担当者までお問い合わせください。

インテル® テクノロジーの機能と利点はシステム構成によって異なり、対応するハードウェアやソフトウェア、またはサービスの有効化が必要となる場合があります。詳細については、http://www.intel.co.jp/ をご覧いただくか、ハードウェア・メーカーまたは販売店にお問い合わせください。絶対的なセキュリティーを提供できるコンピューター・システムはありません。

テストは、特定のシステムでの個々のテストにおけるコンポーネントのパフォーマンスを実証します。ハードウェア、ソフトウェア、システム構成などの違いにより、実際の性能は掲載された性能テストや評価とは異なる場合があります。購入を検討される場合は、ほかの情報も参考にして、パフォーマンスを総合的に評価することをお勧めします。性能やベンチマーク結果について、さらに詳しい情報をお知りになりたい場合は、http://www.intel.com/performance/ (英語) を参照してください。

記載されているコスト削減シナリオは、指定の状況と構成で、特定のインテル® プロセッサー搭載製品が今後のコストに及ぼす影響と、その製品によって実現される可能性のあるコスト削減の例を示すことを目的としています。 状況はさまざまであると考えられます。 インテルは、いかなるコストもコスト削減も保証いたしません。

四半期、年度、および将来の計画と予想について言及している本資料内の記述は、多数のリスクや不確定要素を伴う将来の見通しです。インテルの業績および計画に影響を及ぼす可能性のある要素の詳細については、Form 10-K の年次報告書を含む、インテルの SEC 提出資料に記載されています。

本資料で説明されている製品には、エラッタと呼ばれる設計上の不具合が含まれている可能性があり、公表されている仕様とは異なる動作をする場合があります。現在確認済みのエラッタについては、インテルまでお問い合わせください。

本資料は、(明示されているか否かにかかわらず、また禁反言によるとよらずにかかわらず) いかなる知的財産権のライセンスも許諾するものではありません。

インテルは、本資料で参照しているサードパーティーのベンチマーク・データまたはウェブサイトについて管理や監査を行っていません。本資料で参照しているウェブサイトにアクセスし、本資料で参照しているデータが正確かどうかを確認してください。

Intel、インテル、Intel ロゴ、Intel Inside ロゴ、Intel Atom、Intel Core、Xeon、Movidius、Intel Nervana、Intel Optane、Iris、nGraph、OpenVino、Stratix は、アメリカ合衆国および / またはその他の国における Intel Corporation またはその子会社の商標です。
*その他の社名、製品名などは、一般に各社の表示、商標または登録商標です。

こんな経験ないですか？

ディープラーニング
やりたい！

でも、
アクセラレータ
（GPGPUとか）
を買う余裕はない。

うーん。。。

クライアント用

まずは普段使っているCPUを使ってみませんか？

サーバー用

CPUの特徴

あらゆるワークロードに対応できる汎用性と柔軟性がCPUの特徴です。

ディープラーニング高速化の要　～AVX命令～

インテルCPUにはAVXというSIMD命令が搭載されており、ディープラーニングにて多用される積和演算に非常に有効です。

Intel® AVX

(Intel® Advanced Vector Extensions)

※最新版は“AVX-512“

はいってる

第4世代Haswellから搭載
第10世代Ice LakeからAVX-512搭載

はいってる

Sandy Bridge世代から搭載
Skylake世代からAVX-512搭載

インテルのAI関連ソフトウェア

オープン AI ソフトウェアの活用

- お手軽にAIアプリを作る
- DLモデルを高速推論する
- Pythonからも使えます！

ツールキット
アプリケーション開発者

ディープラーニング

OpenVINO™ ツールキットの
インテル® ディストリビューション[1]
Caffe、TensorFlow*、MXNet*、ONNX*、Kaldi 用に
CPU/GPU/FPGA/VPU に
ディープラーニング推論を導入

スケー

分散型ディープラー

ライブラリー
データ・サイエンティスト

マシンラーニング (ML)

Python*
- scikit-learn
- Pandas
- NumPy

R
- Cart
- Random Forest
- e1071

分散型
- MLLib (Spark* 上)
- Mahout*

ディープラーニング・フレームワーク

CPU などに最適化

TensorFlow | mxnet | Caffe2 PyTorch | BigDL FOR Spark

その他のフレームワーク最適化が進行中
(PaddlePaddle*、CNTK* など)

ステータス & インストール・ガイド

カーネル
ライブラリー開発者

アナリティクスと ML

Python*向け インテル® ディストリビューション
マシンラーニングに最適化されたインテル® ディストリビューション

インテル® データ・アナリティクス・ライブラリー
インテル® データ・アナリティクス・アクセラレーション・ライブラリー
(マシンラーニングを含む)

ディープラーニング

DNNL
(旧称：インテル® MKL-DNN)

CPU / 内蔵グラフィックス向けの
オープンソースの DNN 関数

ディープラーニング・グラフ・コンパイラー

インテル® nGraph™コンパイラー (ベータ版)
複数のフレームワーク (TF、MXNet*、ONNX*) から複数のデバイス (CPU、GPU、NNP) に最適化されたディープラーニング・モデルの演算処理用のオープンソース・コンパイラー

[1]オープンソース・バージョンは、https://01.org/openvinotoolkit で 入手可能。 *その他の社名、製品名などは、一般に各社の表示、商標または登録商標です。
上記の開発者ペルソナは各行の主なユーザー層を表していますが、横断的な利用を排除するものではありません。
すべての製品、コンピューター・システム、日付、および数値は、現在の予想に基づくものであり、予告なく変更されることがあります。

OpenVINO™ ツールキット

https://software.intel.com/en-us/openvino-toolkit

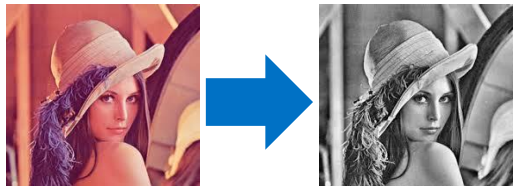
画像処理とディープラーニング推論のためのライブラリスイートです。3つの特徴をぜひご理解ください。

OpenVINO™

コンピュータビジョンアプリ向け
ソフトウェア・ライブラリ・スイート
（Python、C++対応）

Ubuntu,
CentOS,
Yocto,
Win10
MacOS

画像処理



ディープ
ラーニング推論

【特徴１】AIパーツ

【特徴２】モデルコンパイラ

【特徴３】
ヘテロジニアス・オーケストレータ

特徴１．AIパーツとしてのOpenVINO

学習済みモデルとサンプルを多数提供しております。アプリケーション導入の迅速化を促進します。

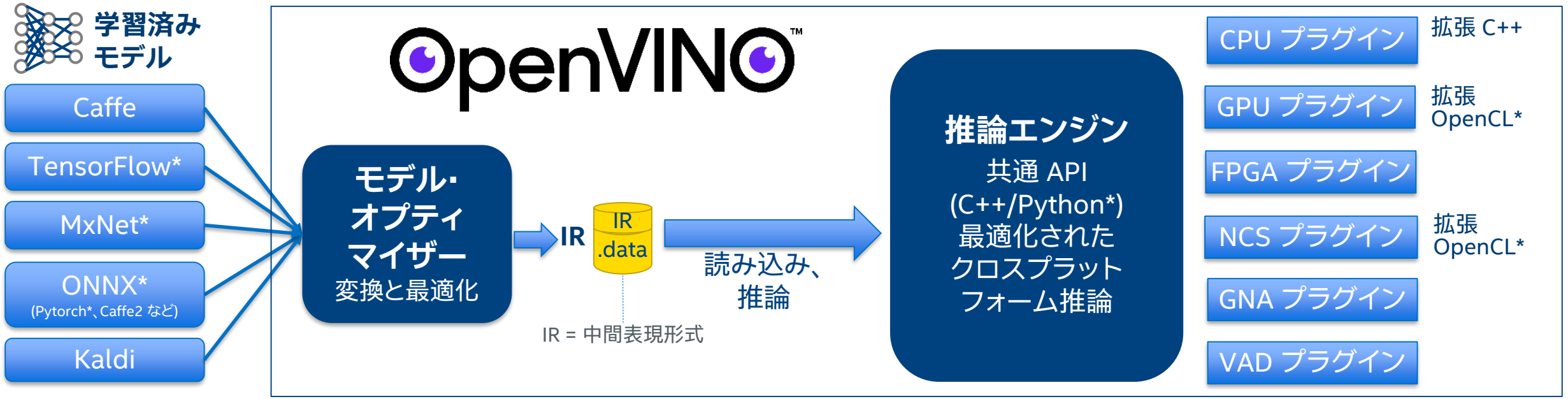
OpenVINO™ ツールキットのインテル® ディストリビューション提供の事前学習済みモデル

- 年齢と性別
- 顔検出 - 標準および拡張
- 頭の位置
- 人物検出 - 眼高 / 高角検出
- 人、車、自転車の検出
- ナンバープレート検出: 小型および前面
- 車両メタデータ
- 人体姿勢推定
- 行動認識 - エンコーダーとデコーダー
- テキストの検出と認識
- 車両検出
- 小売環境
- 歩行者検出
- 歩行者と車両の検出
- 横断者の属性認識
- 感情認識
- 特定の人物をさまざまな動画で識別 - 標準および拡張
- 顔のランドマーク検出
- 視線推定
- 路側物の識別
- 高度な路側識別
- 人の検出と行動認識
- 人の再識別 - 極小 / 超高速
- 顔の再識別
- ランドマーク回帰
- スマート・クラスルームのユースケース
- 単一画像超解像 (3 モデル)
- インスタンス・セグメンテーション
- など...

バイナリーモデル

- 顔検出バイナリー
- 歩行者検出バイナリー
- 車両検出バイナリー
- ResNet50 バイナリー

特徴２：モデルコンパイラとしてのOpenVINO



モデル・オプティマイザー

- **概要**: 学習済みモデルをインポートし、中間表現に変換する Python* ベースのツール
- **重要な理由**: トポロジー変換に基づく抑制により、ハードウェアに適したデータ型に変換することで、パフォーマンスを最大化。

推論エンジン

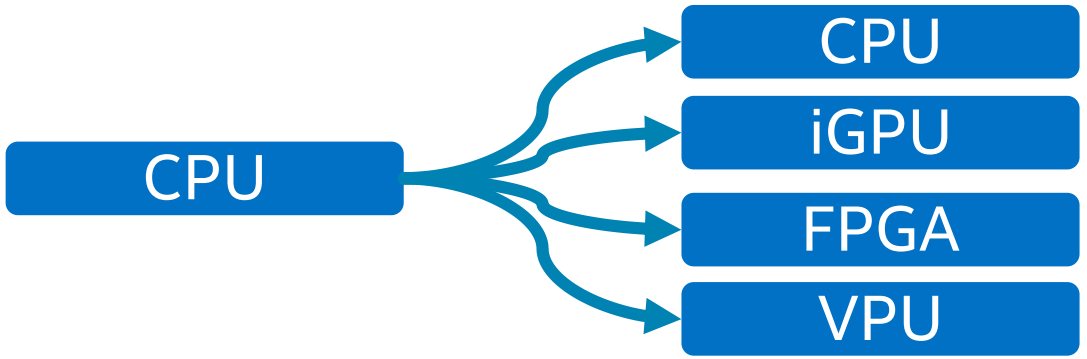
- **概要**: 高レベルの推論 API
- **重要な理由**: インターフェイスは、ハードウェアのタイプに応じた動的読み込みのプラグインとして実装。複数のコードを実装および管理することなく、タイプごとに最適なパフォーマンスを実現可能。

GPU = グラフィックス・プロセシング・ユニット / インテル® プロセッサー・グラフィックスが統合されたインテル® CPU

VAD = ビジョン・アクセラレーター・デザイン・プロダクト。FPGA バージョンと 8 つの Myriad™ X バージョンを含む

特徴３：ヘテロジニアス・オーケストレータとしてのOpenVINO

複数種類のチップが混在するヘテロジニアス環境においても容易に各チップを切り替え、または、併用可能です。

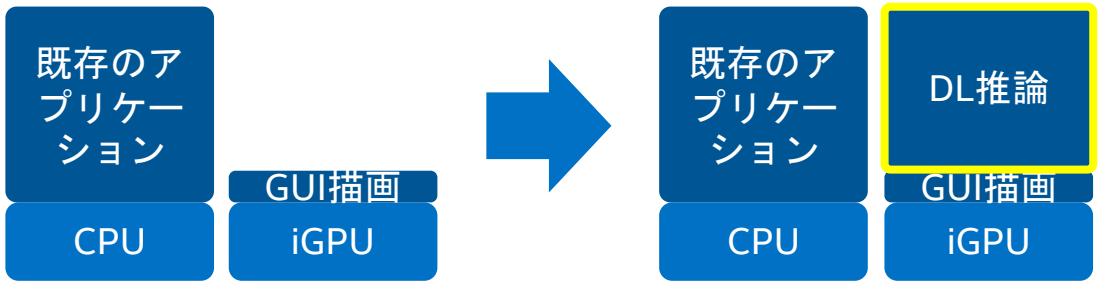


手元のPCで評価、開発着手

必要パフォーマンスが分かった時点でデバイス切り替え

- 厳密な負荷見積もり、ハードウエア選定なしに、すぐに評価、開発着手可能

CPUの処理をiGPUへオフロード



複数チップ併用時にロードバランシング

まとめ　～なぜOpenVINOを使うのか？～

CPUを始めとする各種インテルチップをAI推論環境として使いこなすためです。

OpenVINO - 3つの特徴	期待される効果
【特徴１】AIパーツ	AIアプリケーションの開発効率向上
【特徴２】モデルコンパイラ	インテルチップ（CPU、iGPU、VPUなど）上での推論性能向上
【特徴３】 ヘテロジニアス・オーケストレータ	複数インテルチップの利用効率向上

その他の注目情報

- 最新バージョン
  - 2019R3がリリース！
- 各種操作のGUI化
  - Deep Learning Workbenchの登場！
- 画像認識以外のモデルへも対応
  - LSTM、GMNT、BERTなど自然言語認識系モデルへも対応！

AI の理論と実装の間のハードルを克服

インテルとの連携により AI の取り組みが加速

データレイヤーに関する
インテルの専門知識を活用して
データの氾濫を制御

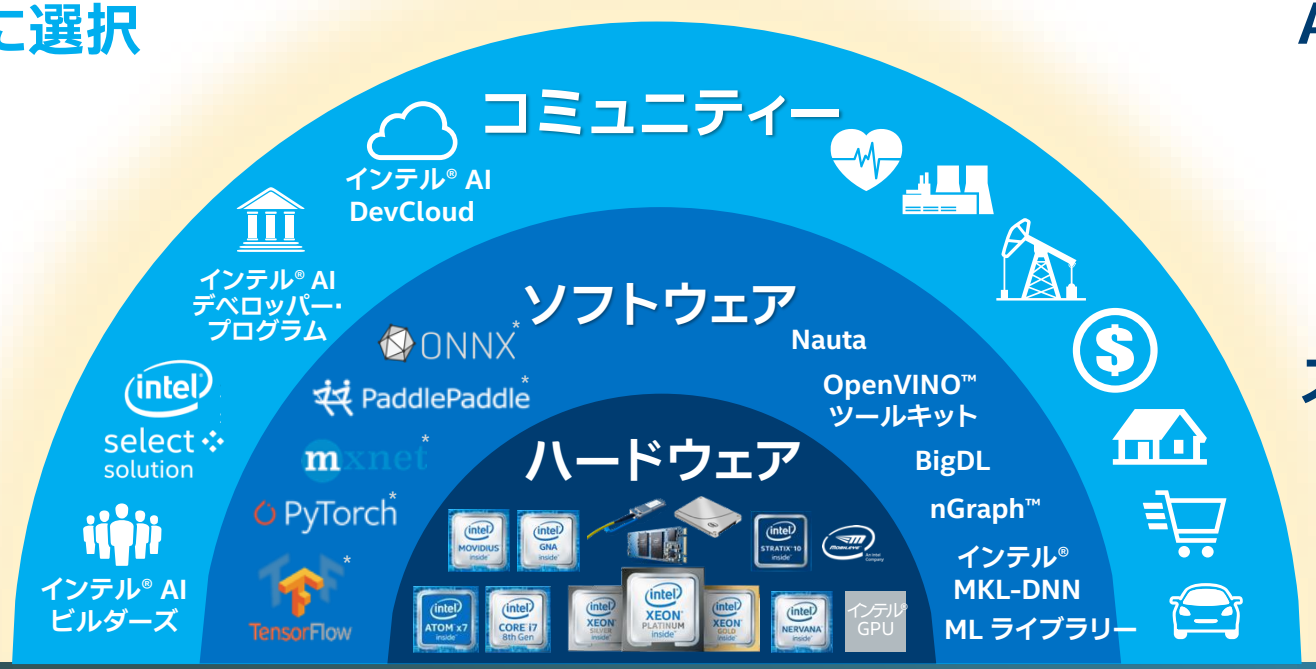
オープン AI ソフトウェアにより
開発を迅速化

アナリティクスから
ディープラーニングまで
アプローチを柔軟に選択

先進的なハードウェアの採用により
AI を広範に展開

専門性の高いコミュニティー
による支援を通じて
AI を簡素化

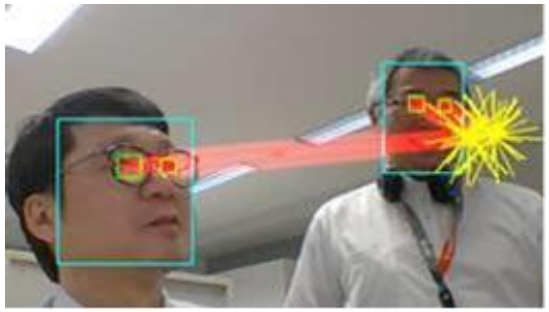
IT 環境およびクラウドの
プラットフォーム全体で
スムーズにスケーリング



*その他の社名、製品名などは、一般に各社の表示、商標または登録商標です。

http://www.intel.ai/ (英語)

デモのソースコード公開中！！



https://github.com/hiouchiy/OpenVINO_Sample/

ありがとうございました