ディープラーニング向け インテル® ソフトウェア製品の紹介

intel Software

久保寺　陽子

AI
人工知能

高度なプログラミングをすることなく、人間の心理と関連した認識機能を実行するために、経験を基にした機械の学習する能力

最適化に関する注意事項
© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。

マシンラーニング技術の分析

訓練

モデル構築のための訓練

- モデル構築時間は重要

推論

モデルを分類、予測、認識に適用

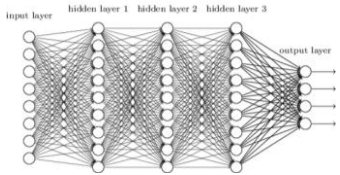
- 簡単に分散処理
- 基準: スループット、TCO @ scale

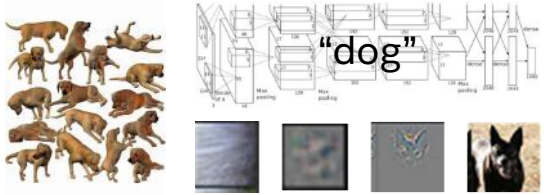
マシンラーニング

経験 (データ) から自主的に学ぶ計算手法

ディープラーニング

多くの隠れたレイヤーへの階層的アプローチ
– 画像や音声、自然言語などの正確に分類されたデータからフレームを得る。特徴は学習される。

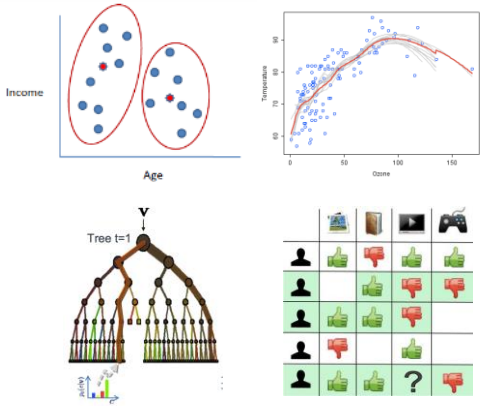




代表的なユーザー: CSP、HPC

その他 (従来) の ML

極めて少ない (1 つまたは 2 つの) 隠れたレイヤーを使ってクラスタリング、回帰、分類を行う従来のマシンラーニングの手法。特徴エンジニアリングが必要。



代表的なユーザー: Enterprise、HPC

最適化に関する注意事項
© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。

インテルのコンピューター・ビジョン・ポートフォリオ

ソリューション

ツール

Intel® Media SDK/Media Server Studio
OpenVINO
Intel® System Studio
Intel® SDK for OpenCL™ Applications
Intel® Parallel Studio XE
oneAPI

フレームワーク

Spark MLlib BigDL
TensorFlow
ONNX
mxnet
PYTORCH
Caffe

ライブラリー

Intel® Data Analytics Acceleration Library
Intel® Distribution for python
Intel® Math Kernel Library
Intel® Nervana™ Graph
Movidius Stack

ハードウェア

Compute
Memory & Storage
Networking
Visual Intelligence

UNLEASH FULL POTENTIAL

最適化に関する注意事項
© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。

インテル® ディープラーニング・ブースト

インテル® Xeon® スケーラブル・プロセッサー (開発コード名 Cascade Lake)

インテル® Xeon® スケーラブル・プロセッサーでは数値精度の低い命令を使って組み込まれたディープラーニングのアクセラレーションが可能

推論

Q3' 2017-Q3' 2018

FP32 精度

Q3' 2018-Q4' 2018

INT8 精度

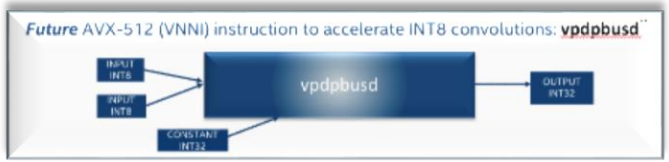
インテル® Xeon® スケーラブル・プロセッサー (開発コード名 Skylake)

Q4' 2018- 2019

VNNI サポート INT8

インテル® Xeon® スケーラブル・プロセッサー (開発コード名 Cascade Lake)

インテル® Xeon® スケーラブル・プロセッサー



インテル® ディープラーニング・ブースト対応 インテル® Xeon® スケーラブル・プロセッサー

https://www.intel.com/content/www/us/en/now/your-data-on-intel/deep-learning-boost-video.html (英語)
https://software.intel.com/en-us/articles/lower-numerical-precision-deep-learning-inference-and-training (英語)

最適化に関する注意事項
© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。

VNNI

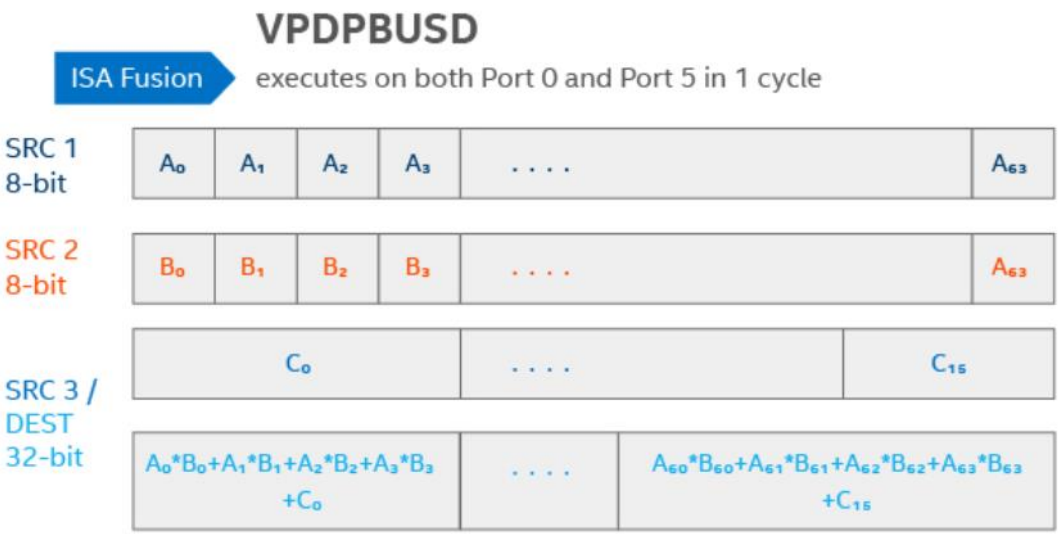
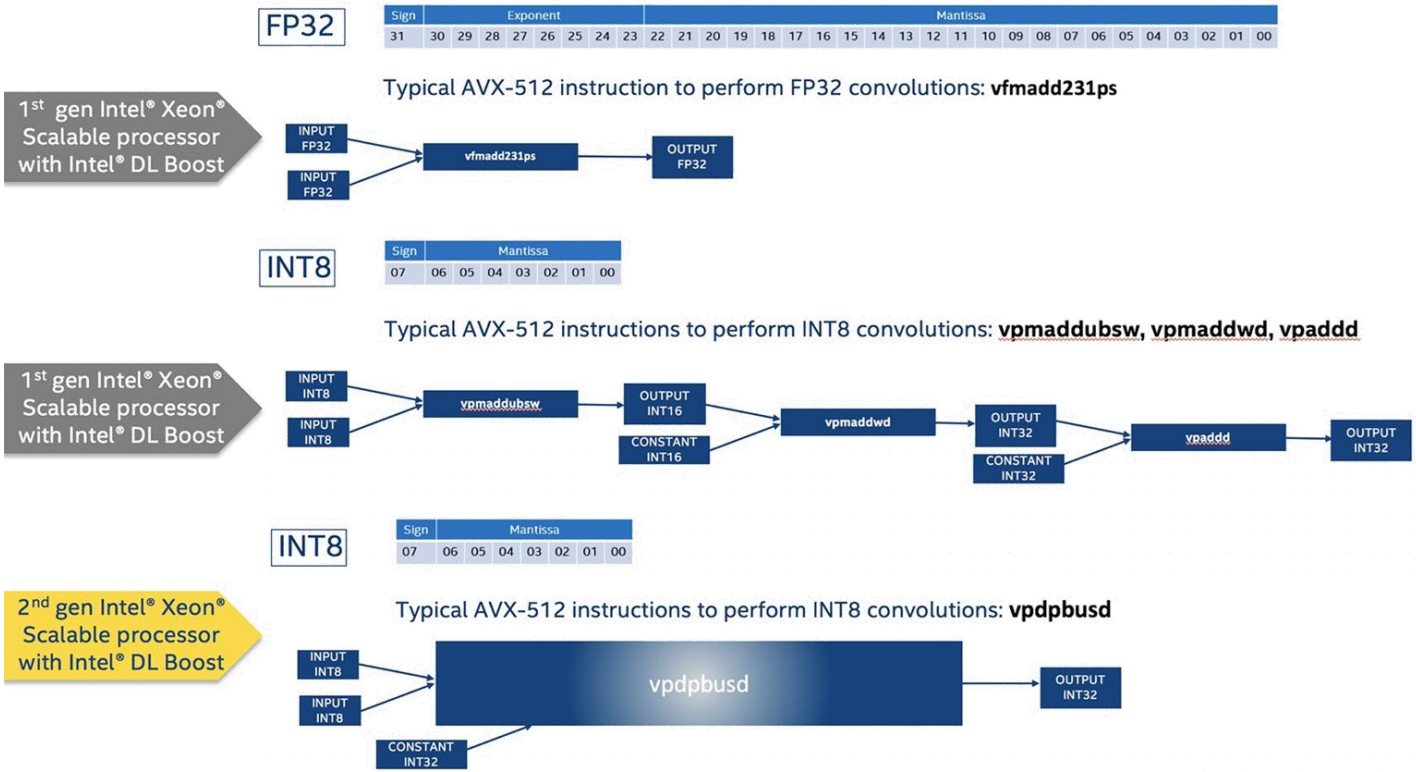


Figure 3. *The Intel DL Boost AVX512_VNNI VPDPBUSD instruction enables 8-bit multiplies with 32-bit accumulates with 1 instruction u8×s8→s32 providing a theoretical peak compute gain of 4x int8 OPS over fp32 OPS. Image credit to Israel Hirsh.*

最適化に関する注意事項
© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。

VNNI



最適化に関する注意事項
© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。

Python* 向けインテル® ディストリビューション 2020

https://software.intel.com/content/www/us/en/develop/tools/distribution-for-python.html (英語)

Python* 向けインテル® ディストリビューションで Python* を高速化

科学計算、データ解析、マシンラーニング、ディープラーニング向けのハイパフォーマンスな Python*

高速なパフォーマンス

パフォーマンス・ライブラリー、並列化、マルチスレッド化、言語拡張

- インテル® MKL[1] とインテル® oneDAL[2] により NumPy*/SciPy*/scikit-learn を高速化
- scikit-learn、daal4py によるデータ解析、マシンラーニング、ディープラーニング
- Numba と Cython によるスケーリング
- 最適化された mpi4py、Dask & PySpark で動作
- 最新のインテル® アーキテクチャー向けに最適化

優れた生産性

事前ビルドの高速化されたパッケージ

- 数値計算、マシンラーニング/ディープラーニング、HPC、データ解析向けの事前ビルドの最適化されたパッケージ
- 既存の Python* から簡単に移行可能 - コード変更不要
- Matplotlib を含む
- 商用開発を含むすべてのユーザーに無料で提供

エコシステムの互換性

Python* 3.7、conda、pip をサポート

- Python* 3.7 をサポート、最適化を Anaconda* ディストリビューションに統合
- ディストリビューションおよび最適化されたパッケージは conda、pip、apt get、yum、Docker* Hub から利用可能、さまざまなパフォーマンスの最適化を Anaconda* ディストリビューションに統合
- 最適化はメインの Python* トランクに反映される

オペレーティング・システム: Windows*、Linux*、macOS*[3]

インテル® アーキテクチャー・ベースのプラットフォーム

詳細: xlsoft.com/jp/products/intel/python/

[1] インテル® マス・カーネル・ライブラリー
[2] インテル® oneAPI データ・アナリティクス・ライブラリー
[3] インテル® Parallel Studio XE Composer Edition でのみ利用可能

最適化に関する注意事項
© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。

Python* 向けインテル® ディストリビューションで Python* を高速化

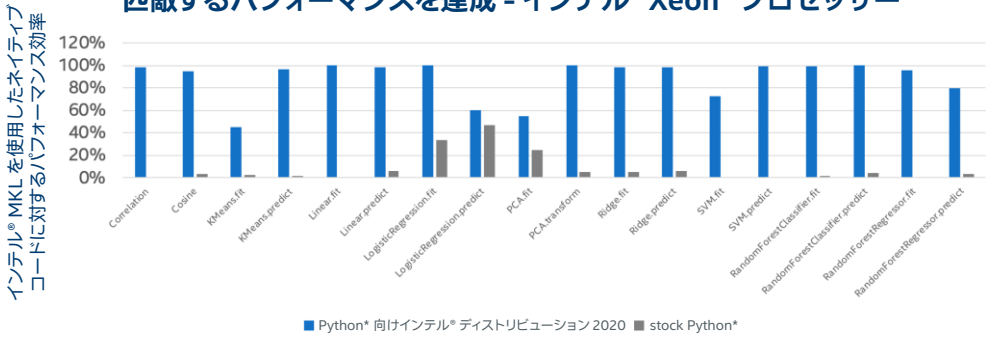
ネイティブコードに近いパフォーマンスを実現

- 科学計算、マシンラーニング、データ解析向けに NumPy*/SciPy*/scikit-learn を高速化
- 既存の Python* から簡単に移行可能 - コード変更不要
- 最新のインテル® プロセッサー向けの高度な最適化

バージョン 2020 の新機能

- Python* のバージョンを 3.7 に更新
- インテル® DAAL を使用して高速化された scikit-learn の主要アルゴリズムによりマシンラーニングのパフォーマンスを向上
- 改良された daal4py パッケージを使用して Python* API でインテル® DAAL 機能を利用するデータ・サイエンティストのニーズに対応

インテルによる Python* 向け最適化はネイティブコード scikit-learn に匹敵するパフォーマンスを達成 - インテル® Xeon® プロセッサー



Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer. Performance results are based on testing as of Nov. 27, 2019 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit www.intel.com/benchmarks.

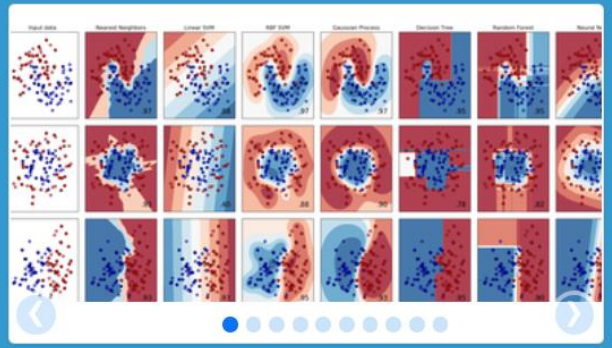
Configuration: Testing by Intel as of Nov. 27, 2019. Stock Python: python 3.7.5 h0371630_0 installed from conda, numpy 1.17.4, numba 0.46.0, llvmlite 0.30.0, scipy 1.3.2, scikit-learn 0.21.3 installed from pip; Intel Python: Intel® Distribution for Python* 2020 Gold: python 3.7.4 hf484d3e_3, numpy 1.17.3 py37ha68da19_4, mkl 2020 intel_133, mkl_fft 1.0.15 py37ha68da19_3, mkl_random 1.1.0 py37ha68da19_0, numba 0.45.1 np117py37_1, llvmlite 0.29.0 py37hf484d3e_9, scipy 1.3.1 py37ha68da19_2, scikit-learn 0.21.3 py37ha68da19_14, daal 2020 intel_133, daal4py 2020 py37ha68da19_4; CentOS Linux 7.3.1611, kernel 3.10.0-514.el7.x86_64; Hardware: Intel(R) Xeon(R) Platinum 8280 CPU @ 2.70GHz (2 sockets, 28 cores/socket, HT: off), 256 GB of DDR4 RAM, 16 DIMMs of 16 GB@2666MHz

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice. Notice revision #20110804

最適化に関する注意事項
© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。

Python* 向けの最も使われるマシンラーニング・パッケージ

https://scikit-learn.org/stable/index.html (英語)



scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying to which category an object belongs to.

Applications: Spam detection, Image recognition.
Algorithms: SVM, nearest neighbors, random forest, …

— Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.
Algorithms: SVR, ridge regression, Lasso, …

— Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, …

— Examples

最適化に関する注意事項
© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。

インテル® DAAL で scikit-learn の最適化

scikit-learn

- マシンラーニングで最も使用されるパッケージ
- 異なるパラメーターで数百ものアルゴリズム
- とても融通が利いて、使いやすいインターフェイス

DAAL4Py

インテル® DAAL は Python* API (ミドルウェア) を備える

インテル® DAAL

インテル® アーキテクチャーにおける高性能な分析とマシンラーニング・アルゴリズム

https://github.com/IntelPython/daal4py (英語)

最適化に関する注意事項
© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。

Python* + ネイティブコードのチューニング

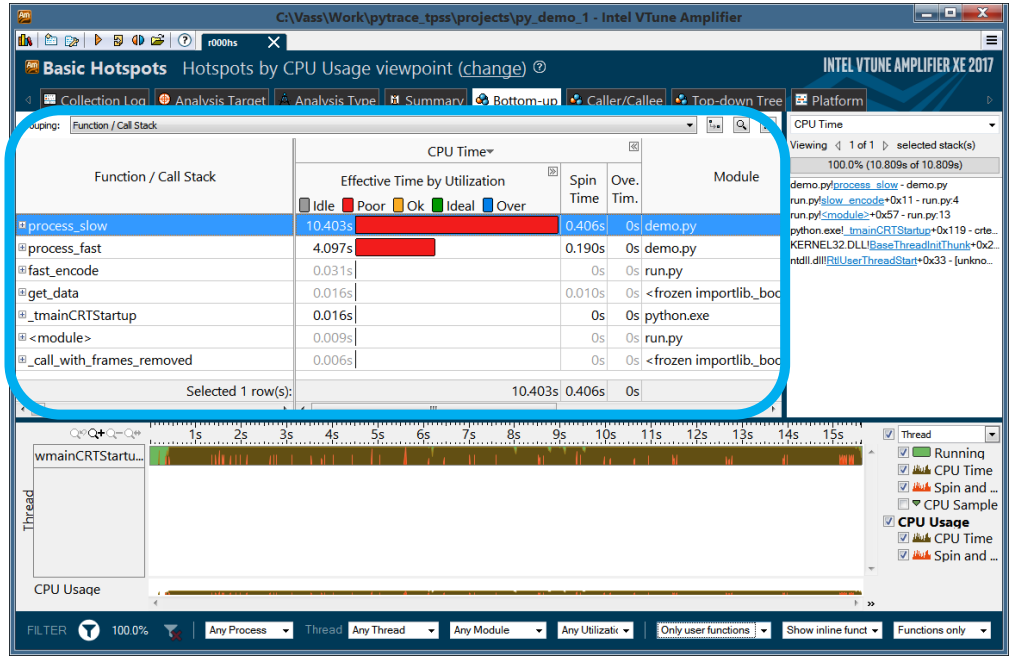
インテル® VTune™ プロファイラーで性能解析 (インテル® Parallel Studio XE に同梱)

課題

- Python* とネイティブコードが混在するアプリケーションを 1 つのツールでプロファイル
- 非効率なランタイム実行を検出

ソリューション

- Python*/C/C++ 混在コードと拡張を自動検出
- パフォーマンス hotspot を行レベルで正確に識別
- 低オーバーヘッド、実行中のアプリケーションへのアタッチ / デタッチ
- パフォーマンスへの影響が大きい個所をチューニング



Python* とネイティブ関数の自動検出およびパフォーマンス解析

インテル® VTune™ プロファイラーとインテル® Parallel Studio XE で可能

最適化に関する注意事項
© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。

Python* 向けインテル® ディストリビューション 2020 のインストール

スタンドアロンのインストーラー

フル・インストーラーのダウンロード
https://software.intel.com/en-us/intel-distribution-for-python (英語)

パッケージ・マネージャーの使用

Anaconda でアクセス
https://software.intel.com/content/www/us/en/develop/articles/using-intel-distribution-for-python-with-anaconda.html (英語)

apt でアクセス
https://software.intel.com/content/www/us/en/develop/articles/installing-intel-free-libs-and-python-apt-repo.html (英語)

Yum でアクセス
https://software.intel.com/content/www/us/en/develop/articles/installing-intel-free-libs-and-python-yum-repo.html(英語)

その他PyPI、Docker でアクセス

3.7

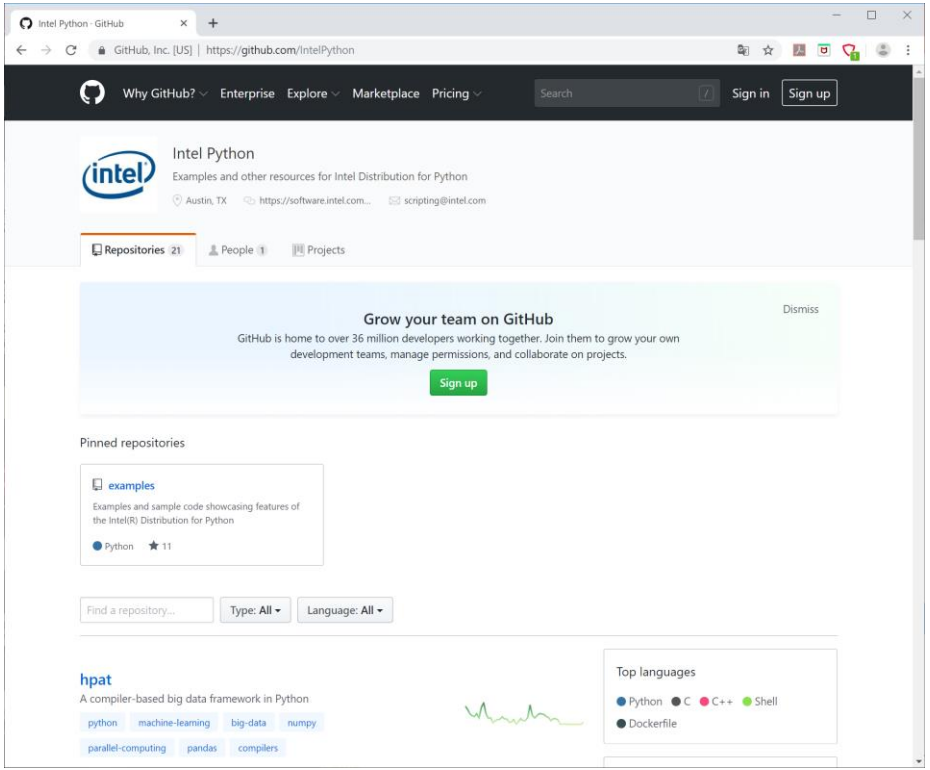
Linux*

Windows*

macOS*

© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。

Python* 向けインテル® ディストリビューションのリソース



https://github.com/IntelPython (英語)

- サンプルプログラムやベンチマークなど
- DAAL4Py も提供

最適化に関する注意事項
© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。

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

(インテル® DAAL)

https://www.xlsoft.com/jp/products/intel/perflib/daal/index.html

インテル® DAAL による解析とマシンラーニングのスピードアップ

マシンラーニングとビッグデータ解析のパフォーマンスを高速化

- アプリケーションがより速くより優れた予測を行えるように支援
- 最高のパフォーマンスを引き出せるようにデータの取り込みとアルゴリズムの計算を最適化
- 広範なアプリケーションのニーズを満たす、オフライン、ストリーミング、分散型使用モデルをサポート
- エッジデバイスとクラウドサービス間で解析ワークロードを分割してアプリケーション全体のスループットを最適化

詳細: xlsoft.com/jp/products/intel/perflib/daal/

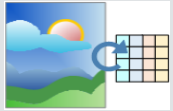
バージョン 2020 の新機能

新しいアルゴリズム

- **ハイパフォーマンス多項 Adaboost**、広く利用されている分類アルゴリズム
- **拡張勾配ブースティング機能**、確率的分類と変数の重要度計算を提供
- **拡張決定木機能**、確率的分類と加重データを提供

前処理 → 変換 → 解析 → モデリング → 検証 → 意思決定

展開、
フィルタリング、正規化

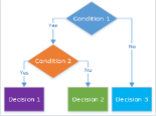


集計、
次元縮小

サマリー統計
クラスタリングなど

マシンラーニング (訓練)、
パラメーター推定、
シミュレーション

仮説検証、
モデルエラー

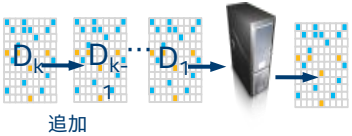


予測、
決定木など

最適化に関する注意事項
© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。

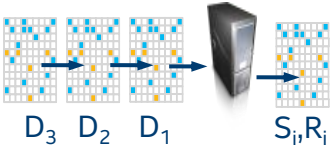
インテル® DAAL の処理モデル

バッチ処理



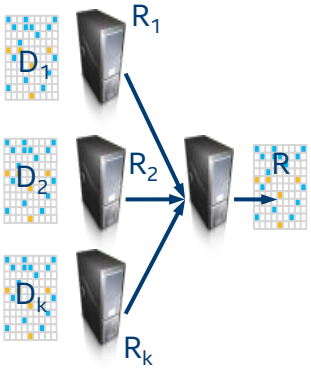
$R = F(D_1,\cdots,D_k)$

オンライン処理



$S_{i+1} = T(S_i,D_i)$
$R_{i+1} = F(S_{i+1})$

分散処理

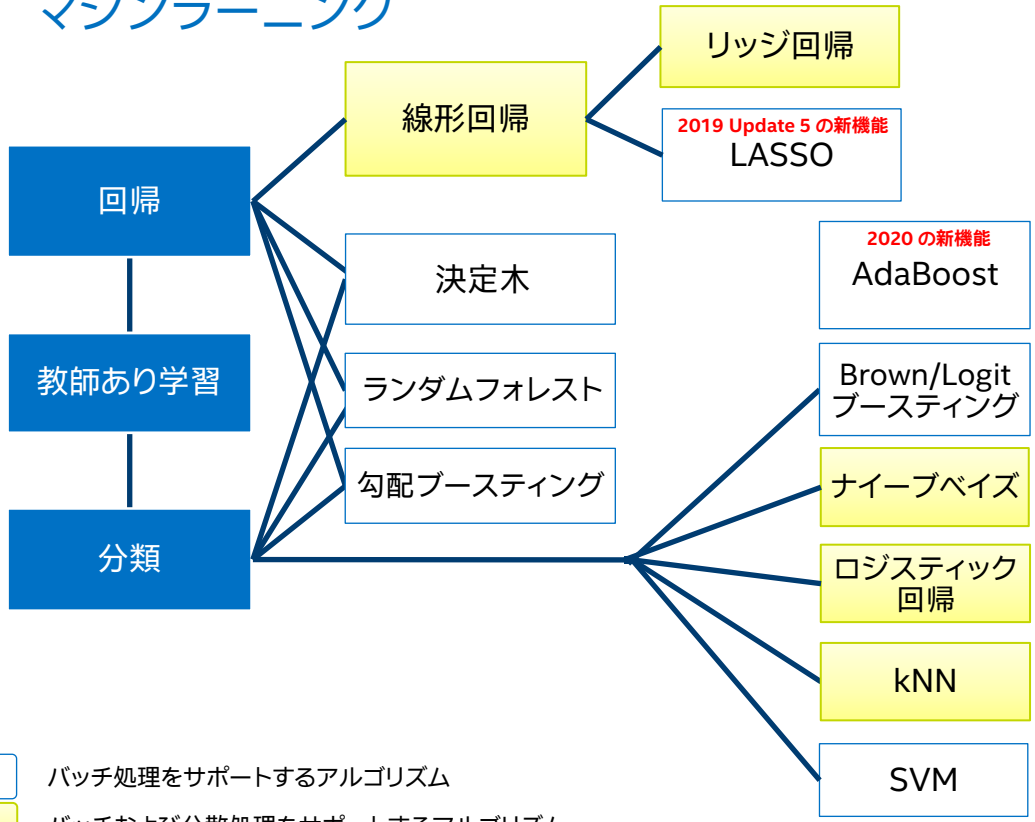


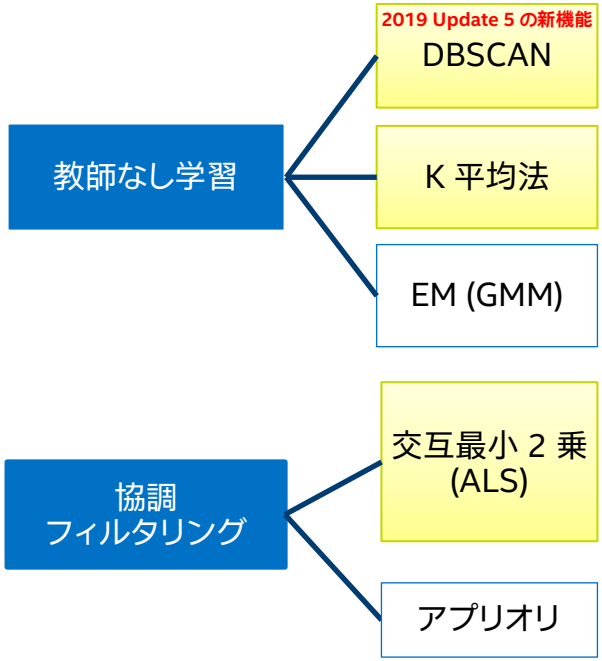
$R = F(R_1,\cdots,R_k)$

最適化に関する注意事項
© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。

インテル® DAAL のアルゴリズム

マシンラーニング





バッチ処理をサポートするアルゴリズム

バッチおよび分散処理をサポートするアルゴリズム

最適化に関する注意事項

© 2020 Intel Corporation. 無断での引用、転載を禁じます。

* 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。

インテル® DAAL のアルゴリズム

データ変換および解析

- データセットの基礎統計
  - 低次モーメント
  - 分位数
  - 順序統計量
- 補正と依存性
  - コサイン距離
  - 相関距離
  - 分散共分散行列
- 行列の因数分解
  - SVD
  - QR
  - コレスキー
  - tSVD
- 次元縮小
  - PCA
  - 相関ルールマイニング(アプリオリ)
  - 最適化ソルバー(SGD、AdaGrad、lBFGS、CD)
- 外れ値検出
  - 単変量
  - 多変量
  - 数学関数(exp、log、…)

バッチ処理をサポートするアルゴリズム

バッチ、オンライン、分散処理をサポートするアルゴリズム

最適化に関する注意事項
© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。

インテル® DAAL による性能のスケール

CPU コア内

- SIMD ベクトル化: 最新の命令セットに最適化、インテル® AVX2、インテル® AVX-512…
- 内部的には、インテル® MKL のシーケンシャル部分を採用

マルチコアやメニーコアへのスケール

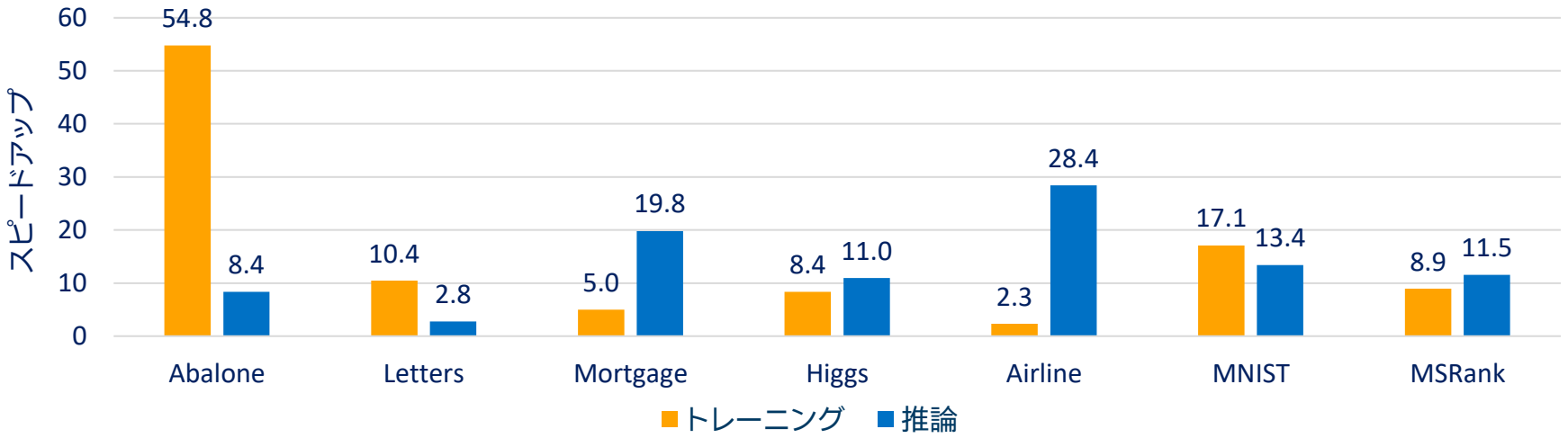
- インテル® スレッディング・ビルディング・ブロックによるスレッド化

クラスターへのスケール

- ユーザーのアプリケーションによる分散処理 (MPI、MapReduce など)
- インテル® DAAL の利用
  - 部分的で中間的な計算結果をサポートするデータ構造
  - 部分的で中間的な計算結果を全体の計算結果にまとめる関数

最適化に関する注意事項
© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。

勾配ブースティングのパフォーマンス (数値が大きいほど高性能)
インテル® DAAL 2020 と DMLC XGBoost 0.9 の比較



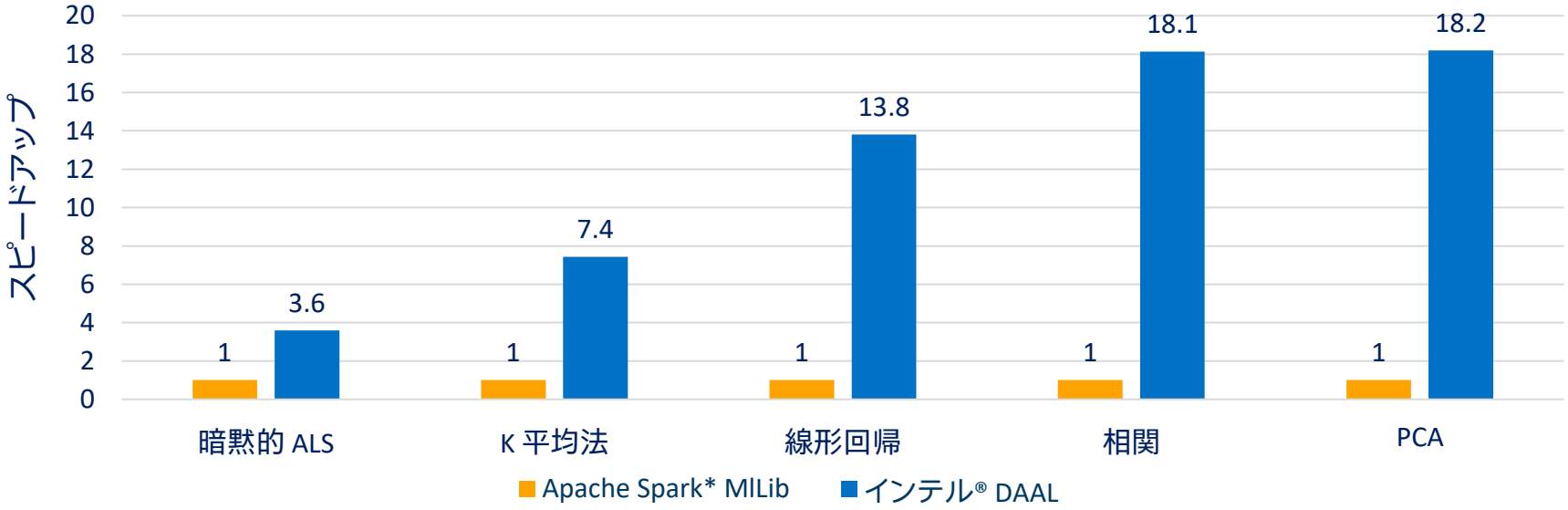
Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer. Performance results are based on testing as of Nov. 11, 2019 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure. Software and workloads used in performance tests may have been optimized. for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

Configuration:. CPU configuration: c5.metal AWS Instance (2nd Generation Intel® Xeon® Scalable Processors 2 sockets, HT:on, Turbo:on, OS: Ubuntu 18.04.2 LTS, Total Memory 193 GB (12 slots/16GB/2933 MHz), BIOS: 1.0 Amazone EC2 (ucode: 0x5000017), OMP Environment: OMP_NUM_THREADS=48 OMP_PLACES={0}:96:1). SW: XGBoost 0.9 download from PIP, Intel DAAL: 2020 version. Python env: Python 3.6, Numpy 1.16.4, Pandas 0.25, Scikit-lean 0.21.2. Parameters for XGBoost: on CPU = { 'alpha': 0.9, 'max_bin': 256, 'scale_pos_weight': 2, 'learning_rate': 0.1, 'subsample': 1, 'reg_lambda': 1, "min_child_weight": 0, 'max_depth': 8, 'max_leaves': 2**8, 'tree_method': 'hist', 'predictor': 'cpu_predictor' }. Parameters for XGB on GPU { 'alpha': 0.9, 'max_bin': 256, 'scale_pos_weight': 2, 'learning_rate': 0.1, 'subsample': 1, 'reg_lambda': 1, "min_child_weight": 0, 'max_depth': 8, 'max_leaves': 2**8, 'tree_method': 'gpu_hist', 'predictor': 'gpu_predictor' }. Input data format: Numpy array . objective functions: 'binary:logistic' for binary classification, 'multi:softmax' for multiclass, 'reg:squarederror' for regressions. Number of iterations: 200 for MNIST, 100 for Mortgage, 1000 for others.

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice. Notice revision #20110804

最適化に関する注意事項
© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。

インテル® DAAL 2020 と Apache Spark* MLlib のパフォーマンス (数値が大きいほど高性能)



Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer. Performance results are based on testing as of Nov. 11, 2019 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure. Software and workloads used in performance tests may have been optimized. for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit www.intel.com/benchmarks.

Configuration: Testing by Intel as of Nov. 11, 2019. 7 x m5.2xlarge AWS instances, Intel® Data Analytics Acceleration Library 2020 (Intel® DAAL); Correlation (# samples = 10M, # features = 1000, (Intel® DAAL=35.2s, MLLib=638.2s)), PCA (# samples = 10M, # features = 1000 (Intel® DAAL=35.2s, MLLib=639.8s)), implicit ALS (# users = 1M, # items = 1M, # factors = 100, # Iterations = 1 (Intel® DAAL=37.6s, MLLib=134.9s)), Linear Regression (# samples = 100M, # features = 50 (Intel® DAAL=16.3s, MLLib=224.5s)), k-means (# samples = 100M, # features = 50, # clusters = 10, # Iterations = 100 (Intel® DAAL=211s, MLLib=1567.3s))

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice. Notice revision #20110804

© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。

推論

OpenVINO™ ツールキット

(Open Visual Inference & Neural Network Optimization)

https://www.xlsoft.com/jp/products/intel/openvino/index.html

OpenVINO™ ツールキットの概要

OpenVINO™ ツールキットはインテル® アーキテクチャーに最適化され
幅広い OS をサポート: Ubuntu*、CentOS*、Yocto*、Windows* 10、macOS*

最適化に関する注意事項
© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。

インテル® ディープラーニング・デプロイメント・ツールキット

インテル® アーキテクチャーの性能を有効活用

モデル・オプティマイザー

- **What it is**: 準備段階 -> トレーニングされたモデルをインポート
- **Why important**: 従来のトポロジー変換で性能/スペースを最適化。ハードウェアがサポートするデータ型への変換では最大の性能向上が見込まれる

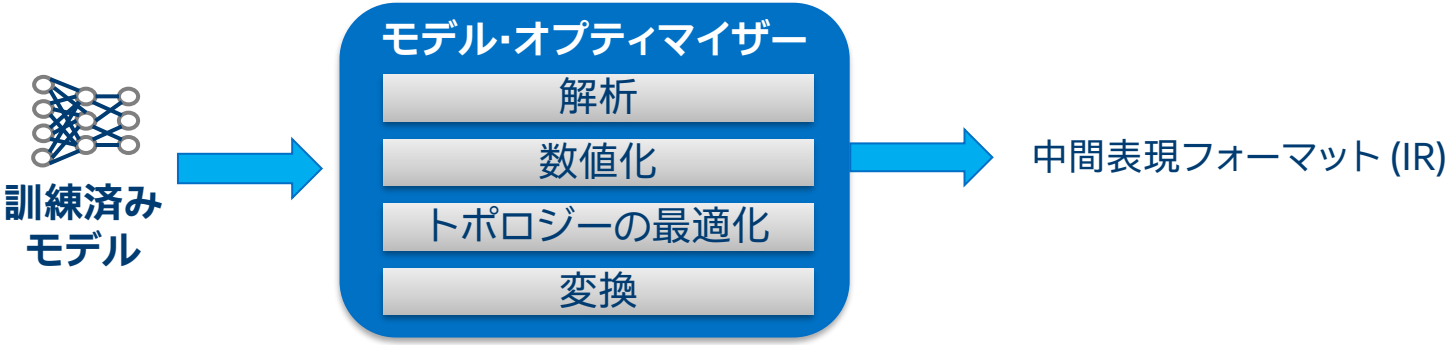
推論エンジン

- **What it is**: 高レベルの推論 API
- **Why important**: インターフェイスは各ハードウェア・タイプごとに直接ロードされるプラグインとして実装。プログラマーが複数のコード経路を実装、維持することなく最高の性能を提供することが可能

GPU = インテグレーテッド・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・グラフィックスを搭載したインテル® CPU

最適化に関する注意事項
© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。
OpenCL および OpenCL ロゴは、Apple Inc. の商標であり、Khronos の使用許諾を受けて使用しています。

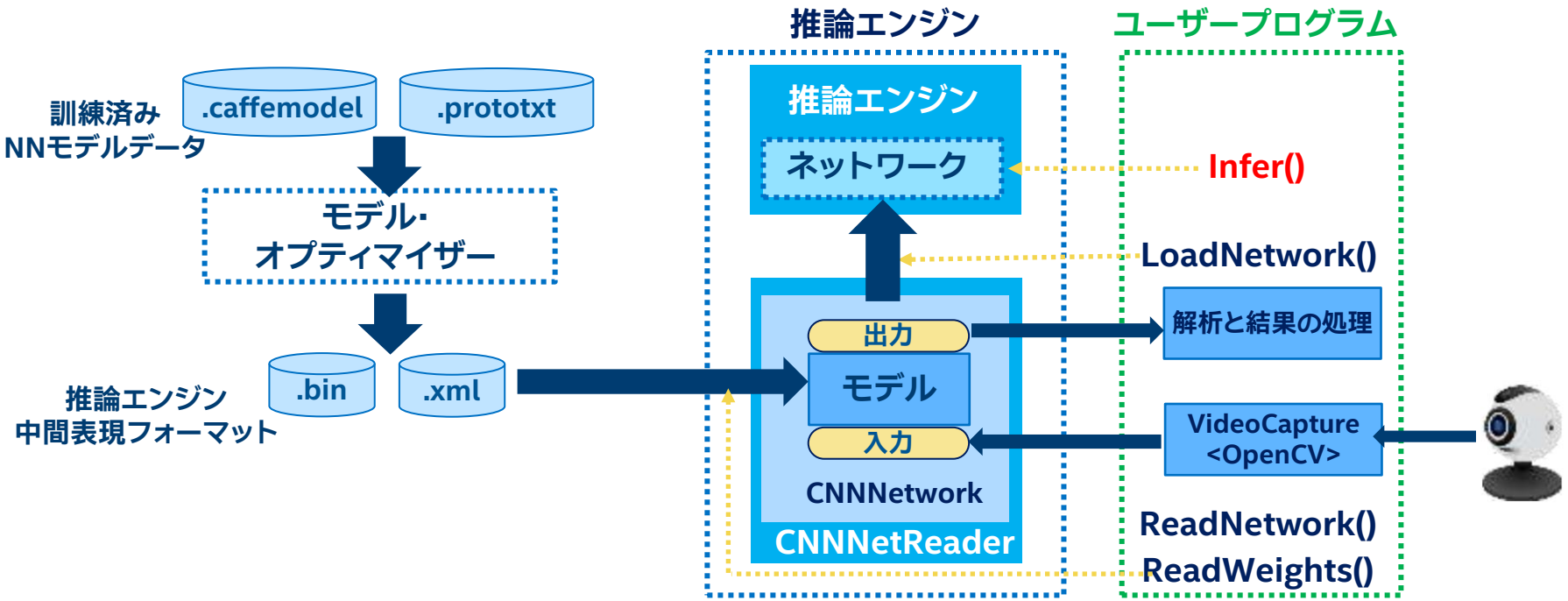
モデル・オプティマイザーによる性能向上



- 使い勝手の良い Python* ベースのワークフローはフレームワークの再構築を要求しない
- サポートされているさまざまなフレームワークからモデルを読み込む - Caffe、TensorFlow*、MXNet*、ONNX*、Kaldi*
- Caffe、MXNet*、TensorFlow* の 100 以上のモデルを検証済み、ONNX* model zoo のすべてのモデルをサポート
- Kaldi* のサポートにより、モデル・オプティマイザーは画像認識以外のネットワークでの推論を拡張
- 中間表現フォーマット (IR) は標準レイヤーを使用しており、ユーザーが提供するカスタムレイヤーは Caffe 不要
- サポートされていないレイヤーの場合には、元のフレームワークにフォールバックが可能だが、オリジナルのフレームワークが必要

最適化に関する注意事項
© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。

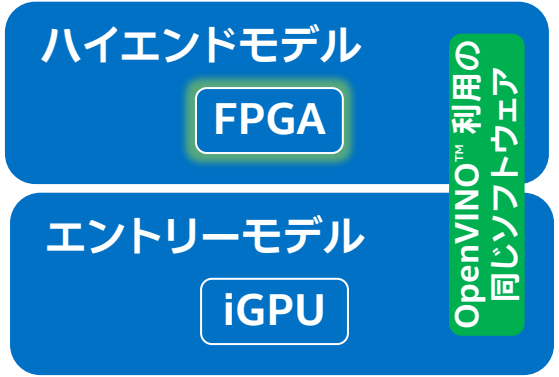
ユーザープログラムからの使用方法

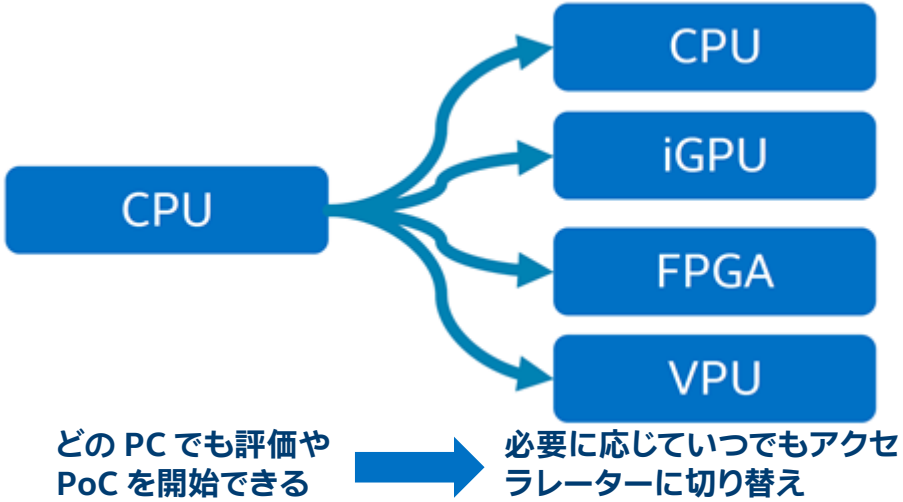


推論エンジンとの利用は簡単!

最適化に関する注意事項
© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。

開発効率を上げる推論エンジンの柔軟性



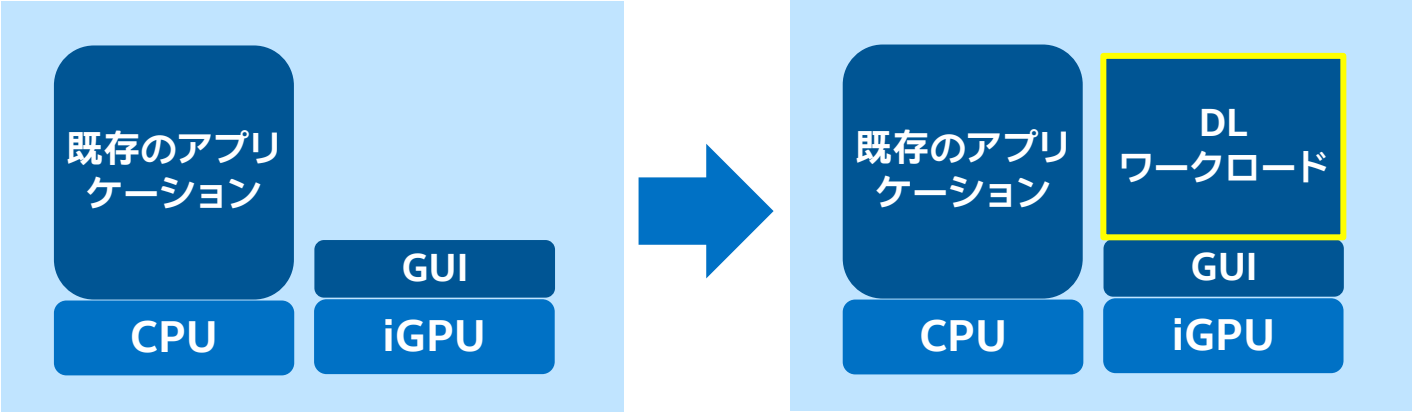


- エントリーからハイエンドまで、異なるハードウェア構成を同じソフトウェアでサポート
- 容易なソフトウェア開発とメンテナンス

- 詳細なワークロードの評価なしでプロジェクトがスタート可能
- 推論エンジンは他のディープラーニング・ソリューションとは異なりハードウェア依存性が少ない

最適化に関する注意事項
© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。

今のデバイスに付加価値を

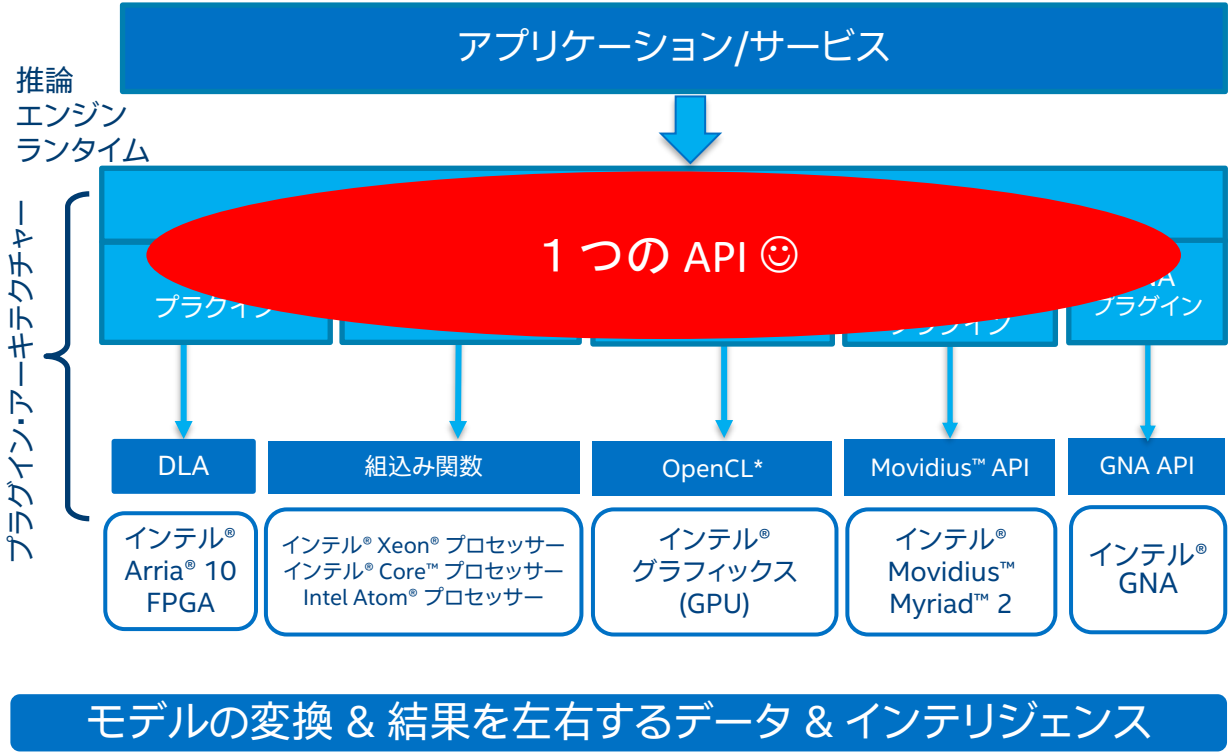


- 推論エンジンは、内蔵 GPU を利用することによって、既存のアプリケーションへの影響を最小限におさえながらユーザーの機能追加を可能に
- 追加費用やデバイスは不要

最適化に関する注意事項
© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。

推論エンジンで最適なモデル性能を実現

- すべてのインテル® アーキテクチャーをサポートする簡単で統一された API
- 幅広いインテル® アーキテクチャー (CPU/GEN/FPGA) をサポートする最適化された推論
- ヘテロジニアスなハードウェア・タイプにまたがったレイヤーの実行をサポート
- 非同期実行により性能を向上
- 次世代のインテル® アーキテクチャー・ベースのプロセッサー向けの大規模開発

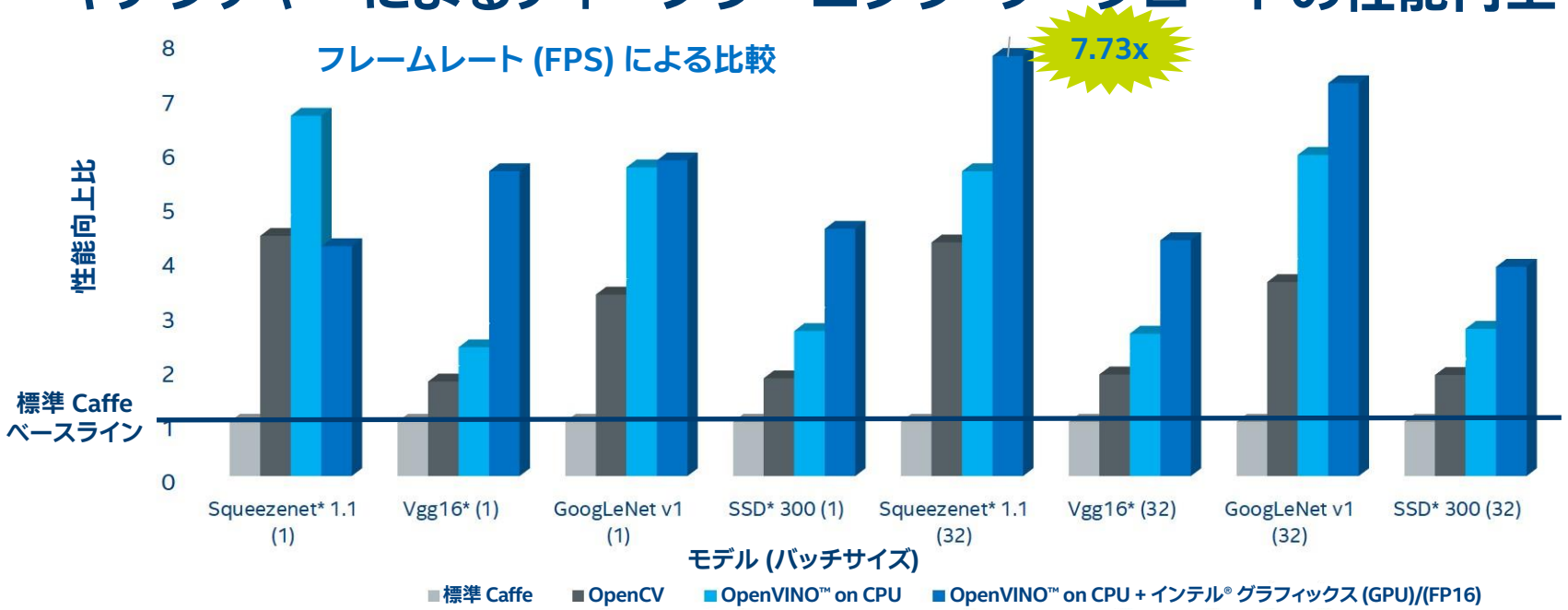


モデルの変換 & 結果を左右するデータ & インテリジェンス

GPU = インテグレーテッド・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・グラフィックスを搭載したインテル® CPU
GNA = ガウス混合モデルおよびニューラル・ネットワーク・アクセラレーター

最適化に関する注意事項
© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。
OpenCL および OpenCL ロゴは、Apple Inc. の商標であり、Khronos の使用許諾を受けて使用しています。

各モデルにおける OpenVINO™ ツールキットとインテル® アーキテクチャーによるディープラーニング・ワークロードの性能向上



インテルのハードウェアではアクセラレーターなしでも高速

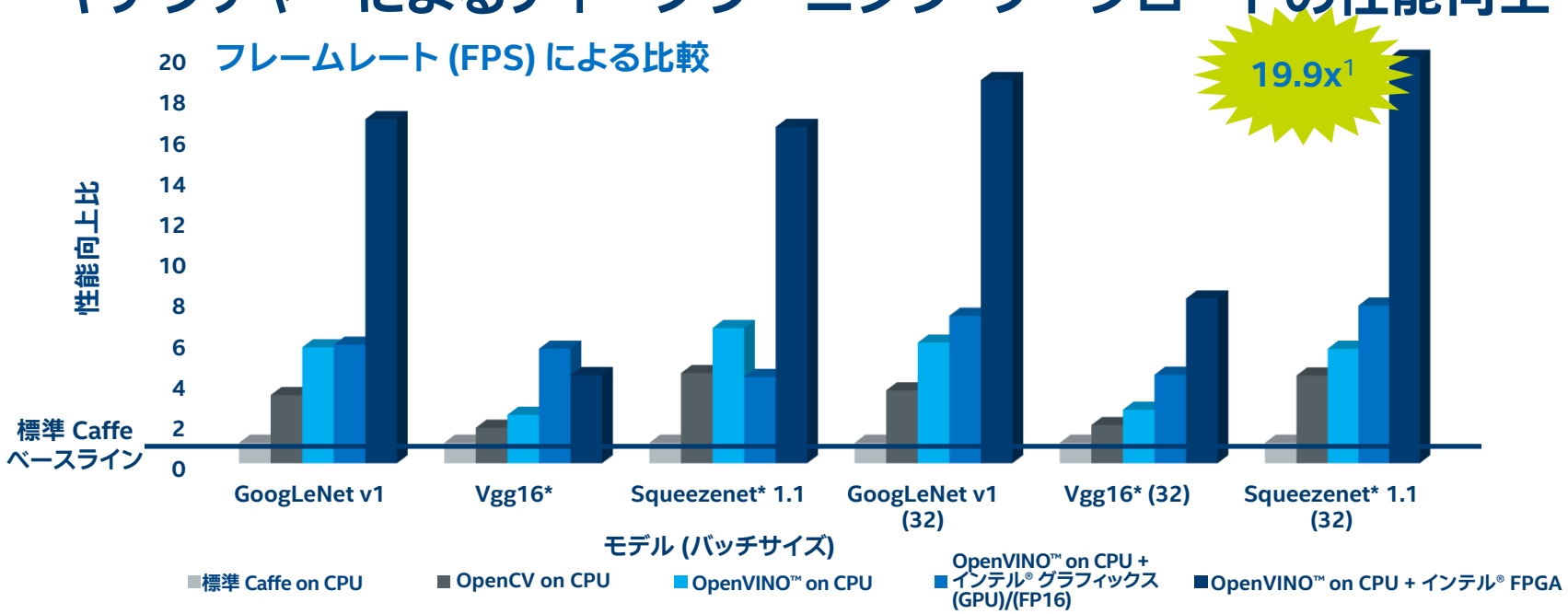
[1]Depending on workload, quality/resolution for FP16 may be marginally impacted. A performance/quality tradeoff from FP32 to FP16 can affect accuracy; customers are encouraged to experiment to find what works best for their situation. The benchmark results reported in this deck may need to be revised as additional testing is conducted. Performance results are based on testing as of April 10, 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. **Configuration:** Testing by Intel as of April 10, 2018. Intel® Core™ i7-6700K CPU @ 2.90GHz fixed, GPU GT2 @ 1.00GHz fixed Internal ONLY testing, Test v312.30 – Ubuntu* 16.04, OpenVINO™ 2018 RC4. Tests were based on various parameters such as model used (these are public), batch size, and other factors. Different models can be accelerated with different Intel hardware solutions, yet use the same Intel software tools.

最適化に関する注意事項

© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this fixed product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice. Notice revision #20110804

各モデルにおける OpenVINO™ ツールキットとインテル® アーキテクチャーによるディープラーニング・ワークロードの性能向上



インテル® FPGA でさらに大きな性能向上

[1]Depending on workload, quality/resolution for FP16 may be marginally impacted. A performance/quality tradeoff from FP32 to FP16 can affect accuracy; customers are encouraged to experiment to find what works best for their situation. Performance results are based on testing as of June 13, 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. **Configuration:** Testing by Intel as of June 13, 2018. Intel® Core™ i7-6700K CPU @ 2.90GHz fixed, GPU GT2 @ 1.00GHz fixed Internal ONLY testing, Test v3.15.21 – Ubuntu* 16.04, OpenVINO 2018 RC4, Intel® Arria® 10 FPGA 1150GX. Tests were based on various parameters such as model used (these are public), batch size, and other factors. Different models can be accelerated with different Intel hardware solutions, yet use the same Intel software tools.

最適化に関する注意事項

© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice. Notice revision #20110804

OpenVINO™ ツールキットのセミナー/ワークショップ

https://www.xlsoft.com/jp/products/intel/tech/seminar/openvino.html
（無料）

- レベル別に分かれたコース設定: 一般のディープラーニング・フレームワークで構築した新たな学習済みモデルから、OpenVINO™ ツールキットによる推論アプリケーションとして実装するまでの流れを学習可能
  - ベーシックコース: ツールの概要を学習し、サンプルコードをベースにデモが動かせるレベル
  - アドバンスド・コース: ベーシックコースの受講後に、独自のモデルを構築・最適化して動かせるレベル
  - ワークショップ: 実際に手を動かしながら順を追ってツールの使い方を学習するためのハンズオン・トレーニングを実施

※ベーシックコース、アドバンスド・コースは座学のみのコース
スライド資料は、セミナー開催後に電子データのダウンロード URL のみを案内する予定

最適化に関する注意事項
© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。

その他のツール

インテル® パフォーマンス・ライブラリーの統合

- ディープ・ニューラル・ネットワーク向けインテル® マス・カーネル・ライブラリー (インテル® MKL-DNN)
  - インテル® MKL-DNN、Deep Neural Network Library (DNNL) → インテル® oneAPI ディープ・ニューラル・ネットワーク・ライブラリー (インテル® oneDNN)
    - 変更点: ライブラリー名 → インテル® oneDNN、リポジトリー “intel/mkl-dnn” → “opeapi-src/oneDNN”
    - API、環境変数、ビルドオプションは現時点で変更なし

https://github.com/oneapi-src/oneDNN (英語)

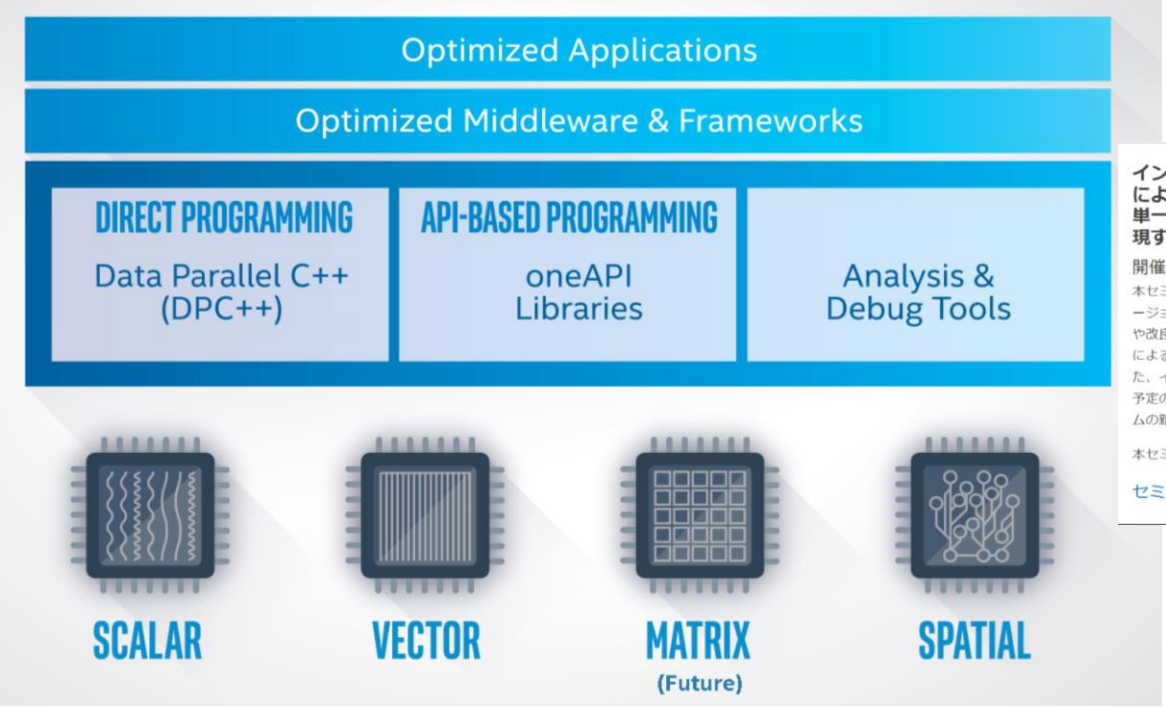
- インテル® Machine Learning Scaling Library (インテル® MLSL)
  - インテル® oneAPI コレクティブ・コミュニケーション・ライブラリー (インテル® oneCCL) に移行
    - 分散訓練フレームワーク Horovod に統合
      https://github.com/horovod/horovod/blob/master/docs/oneccl.md (英語)
    - 機械学習フレームワーク Pytorch* に統合
      https://github.com/intel/torch-ccl (英語)

https://oneapi-src.github.io/oneCCL (英語)

最適化に関する注意事項
© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。

oneAPI とは?

https://software.intel.com/content/www/us/en/develop/tools/oneapi.html (英語)



過去のセミナーでも紹介
https://www.isus.jp/event/

インテル® Parallel Studio XE 2020 によるアプリケーション最適化入門 ～ 単一のソース言語でコードの再利用を実現する新しい開発環境のご紹介 ～

開催日: 2020年 1月 17日 (金)

本セミナーでは、インテル Parallel Studio XE の新バージョン 2020 に同梱されるコンポーネントの新機能や改良点を説明し、さらにパワーアップした本ツールによるパフォーマンスの最適化手法を紹介します。また、インテル社のゲストスピーカーをお招きして開催予定のセッションでは、現在実施中のベータプログラムの新しい開発環境についてもあわせて紹介します。

本セミナーの資料を公開します。

セミナー詳細はこちら »

セミナー資料

- はじめに: インテル® Parallel Studio XE 2020 とインテル® oneAPI
- インテル® Parallel Studio XE 2020 の新機能
- インテル® Optane™ DC パーシステント・メモリーを使用するアプリケーションの最適化
- 生まれ変わったインテル® VTune™ プロファイラー

© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。

インテル® oneAPI ベース・ツールキット

インテル® oneAPI ベース・ツールキット

ダイレクト・プログラミング

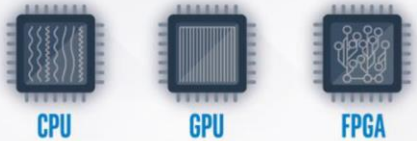
- インテル® oneAPI DPC++ コンパイラー
- インテル® DPC++ 互換性ツール
- Python* 向けインテル® ディストリビューション
- [オプション] oneAPI ベース・ツールキット用インテル® FPGA アドオン

API ベースのプログラミング

- インテル® oneAPI DPC++ ライブラリー
- インテル® oneAPI マス・カーネル・ライブラリー
- インテル® oneAPI データ・アナリティクス・ライブラリー
- インテル® oneAPI スレッディング・ビルディング・ブロック
- インテル® oneAPI ビデオ・プロセシング・ライブラリー
- インテル® oneAPI コレクティブ・コミュニケーション・ライブラリー
- インテル® oneAPI ディープ・ニューラル・ネットワーク・ライブラリー
- インテル® インテグレーテッド・パフォーマンス・プリミティブ

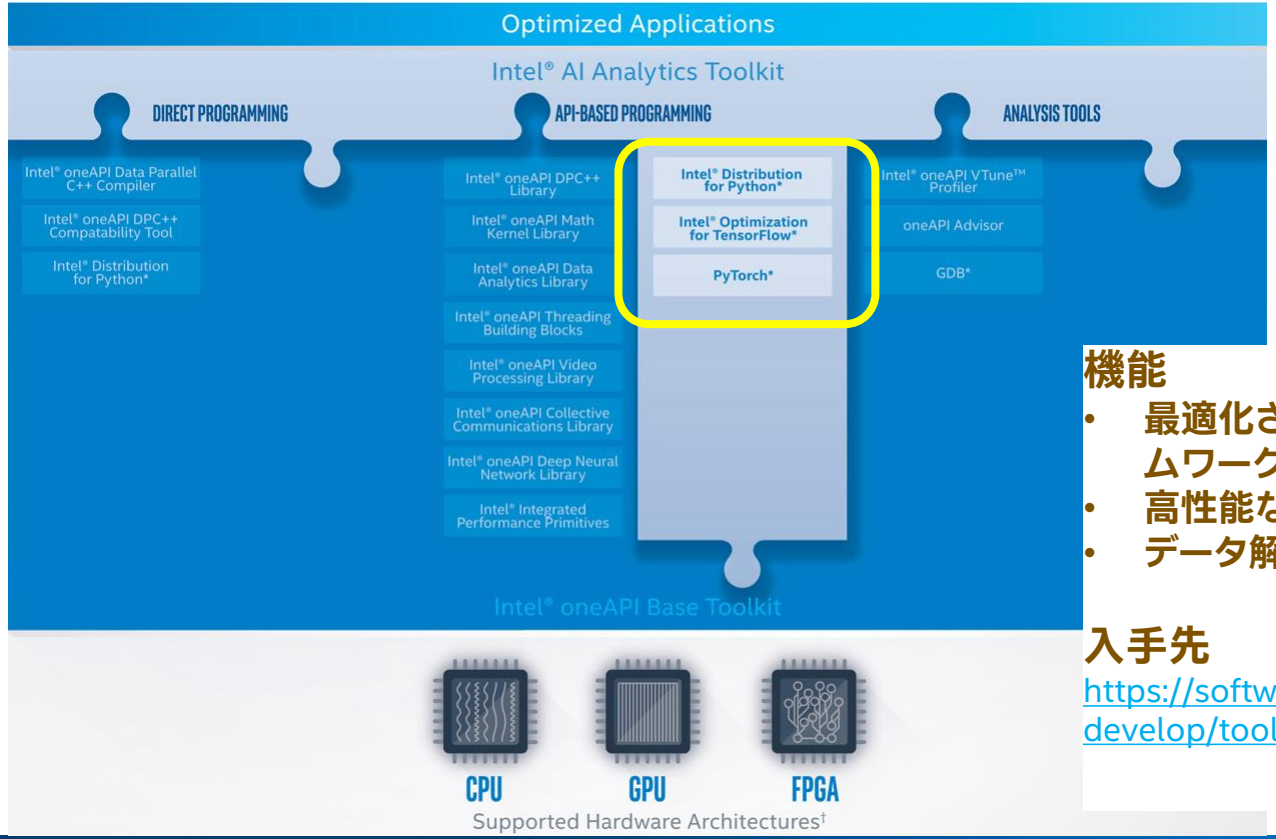
解析ツール

- インテル® VTune™ プロファイラー
- インテル® Advisor
- GDB*



出典: https://www.isus.jp/event/

ベータ版インテル® AI アナリティクス・ツールキット (Linux* のみ)



機能

- 最適化されたディープ・ラーニング・フレームワーク
- 高性能な Python*
- データ解析 & マシンラーニングの強化

入手先

https://software.intel.com/content/www/us/en/develop/tools/oneapi.html (英語)

最適化に関する注意事項
© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。

最適化に役立つ情報

さあ、始めよう
最適化に役立つ情報

iSUS (IA Software User Society)

- インテル® ソフトウェア開発製品の日本語版や日本語ヘルプの提供
- インテル コーポレーションのインテル® ソフトウェア・ネットワークの記事の日本語化

インテル® Parallel Studio XE

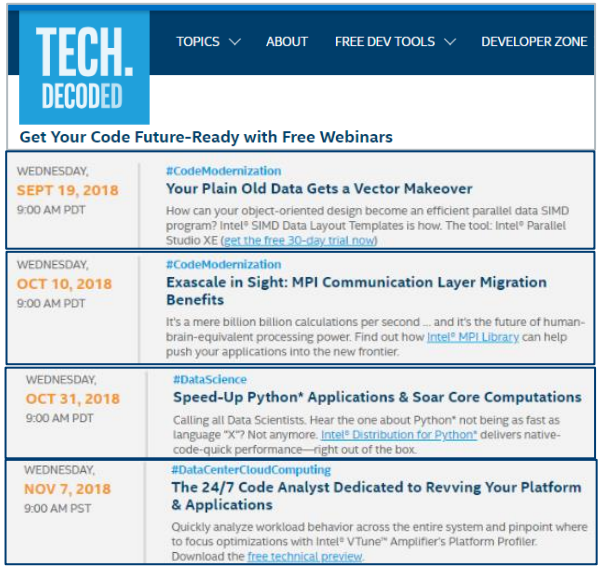
- 概要、機能、サポート、コードサンプル (英語)
- トレーニング資料、Tech.Decoded ウェビナー (英語)、HOW TO ビデオと記事 (英語)
- お客様の声 (英語) & ケーススタディー
- その他のインテル® ソフトウェア開発製品

インテル® Code Modernization Program

- 概要 (英語)
- ライブ・トレーニング (英語)
- リモートアクセス

最適化の近道
TEC ウェビナーを視聴する

登録 (英語)



intel.ly/2PdkNhN (英語)

最適化に関する注意事項
© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。

法務上の注意書きと最適化に関する注意事項

性能に関するテストに使用されるソフトウェアとワークロードは、性能がインテル® マイクロプロセッサー用に最適化されていることがあります。SYSmark* や MobileMark* などの性能テストは、特定のコンピューター・システム、コンポーネント、ソフトウェア、操作、機能に基づいて行ったものです。結果はこれらの要因によって異なります。製品の購入を検討される場合は、他の製品と組み合わせた場合の本製品の性能など、ほかの情報や性能テストも参考にして、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。さらに詳しい情報をお知りになりたい場合は、http://www.intel.com/benchmarks/ (英語) を参照してください。

性能の測定結果はシステム構成の日付時点のテストに基づいています。また、現在公開中のすべてのセキュリティー・アップデートが適用されているとは限りません。詳細については、公開されている構成情報を参照してください。絶対的なセキュリティーを提供できる製品はありません。

インテルは、サードパーティーのデータについて管理や監査を行っていません。原典を確認し、ほかの情報も参考にして、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。

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。実際の性能はシステム構成によって異なります。絶対的なセキュリティーを提供できる製品またはコンポーネントはありません。詳細については、各システムメーカーまたは販売店にお問い合わせいただくか、http://www.intel.co.jp/ を参照してください。

インテルは、明示されているか否かにかかわらず、いかなる保証もいたしません。ここにいう保証には、商品適格性、特定目的への適合性、および非侵害性の黙示の保証、ならびに履行の過程、取引の過程、または取引での使用から生じるあらゆる保証を含みますが、これらに限定されるわけではありません。

Intel、インテル、Intel ロゴ、Intel Inside、Intel Inside ロゴ、Arria、Intel Atom、Intel Core、Movidius、Myriad、Intel Nervana、Xeon、OpenVINO、Intel RealSense、VTune 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。

最適化に関する注意事項

インテル® コンパイラーでは、インテル® マイクロプロセッサーに限定されない最適化に関して、他社製マイクロプロセッサー用に同等の最適化を行えないことがあります。これには、インテル® ストリーミング SIMD 拡張命令 2、インテル® ストリーミング SIMD 拡張命令 3、インテル® ストリーミング SIMD 拡張命令 3 補足命令などの最適化が該当します。インテルは、他社製マイクロプロセッサーに関して、いかなる最適化の利用、機能、または効果も保証いたしません。本製品のマイクロプロセッサー依存の最適化は、インテル® マイクロプロセッサーでの使用を前提としています。インテル® マイクロアーキテクチャーに限定されない最適化のなかにも、インテル® マイクロプロセッサー用のものがあります。この注意事項で言及した命令セットの詳細については、該当する製品のユーザー・リファレンス・ガイドを参照してください。
注意事項の改訂 #20110804

© 2020 Intel Corporation. 無断での引用、転載を禁じます。
* 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。