ソフトウェア AI アクセラレーターで最大 100 倍のパフォーマンス向上を実現

この記事は、インテル® デベロッパー・ゾーンに公開されている「Software AI accelerators: AI performance boost for free」の日本語参考訳です。原文は更新される可能性があります。原文と翻訳文の内容が異なる場合は原文を優先してください。

データの急激な増加は、人工知能 (AI) の旺盛な欲求を満たし、AI はニッチな存在から広く知られる存在へと変貌を遂げました。AI の成長にとって重要なことは、より高い AI パフォーマンスを実現するため、コンピューター・システムの要件が増え続けていることです。CPU、GPU、FPGA などの一般的なチップ・アーキテクチャーに AI アクセラレーションが組み込まれるようになっただけでなく、人工ニューラル・ネットワークやマシンラーニング・アプリケーションの高速化に特化した専用ハードウェア AI アクセラレーターが急増しています。これらのハードウェア・アクセラレーターは目覚ましい AI パフォーマンスの向上をもたらしますが、同じハードウェア構成で、ディープラーニング、古典的なマシンラーニング、グラフ・アナリティクスなどにおいて、桁違いの AI パフォーマンス (英語) を実現するには、ソフトウェア AI アクセラレーターが必要です。さらに、ソフトウェアの最適化によってもたらされるこの AI パフォーマンスの向上は、コードの変更や開発者の時間をほとんど必要とせず、ハードウェアの追加コストも不要であることが特長です。

ソフトウェアによる AI アクセラレーションでもたらされる 10 ~ 100 倍のパフォーマンス向上によって実現できるコスト削減の範囲を可視化してみましょう。例えば、大手ストリーミング・メディア・サービスの多くは、何万時間もの利用可能なコンテンツを持っています。彼らは、コンテンツのモデレーション、テキストの識別、有名人の認識に、画像分類と物体検出アルゴリズムを使用したいと思うかもしれません。また、分類の基準は、地域の習慣や政府の規制など、国によって異なる可能性があり、新しい番組や規則の変更により、毎月コンテンツの約 10% に対してプロセスを繰り返す必要があるかもしれません。大手クラウド・サービス・プロバイダーでこれらの AI アルゴリズムを一定のコストで稼働させた場合、ソフトウェア AI アクセラレーターによってパフォーマンスが 10 倍向上しただけでも、毎月数百万円のコスト削減につながることが分かります**。

同様のコスト削減は、自動キャプション生成や推薦エンジンなどの AI サービスでも実現でき、パフォーマンスが 100 倍向上するケースでは、当然のことながら、さらに高い削減効果が期待できます。AI ワークロードが小さくても、大きなコスト削減効果が期待できる可能性があります。

ソフトウェアが計算プラットフォームの最終的なパフォーマンスを決定するため、ソフトウェア・アクセラレーションは、エンターテインメント、通信、自動車、ヘルスケアなど、さまざまなアプリケーションで「どこでも AI」を実現する鍵となります。

ソフトウェア AI アクセラレーターとは? ハードウェア AI アクセラレーターとの違いは?

ソフトウェア AI アクセラレーターとは、同じハードウェア構成でソフトウェアの最適化によって実現できる AI パフォーマンスの向上を指します。ソフトウェア AI アクセラレーターは、さまざまなアプリケーション、モデル、ユースケースにおいて、プラットフォームを 10 ~ 100 倍以上高速化できます。

AI ワークロードの多様化に伴い、AI 向けに最適化されたさまざまなハードウェア・アーキテクチャーに対するビジネス需要が生まれています。これらは主に、AI アクセラレーション搭載 CPU、AI アクセラレーション搭載 GPU、そして専用ハードウェア AI アクセラレーターの 3 つのカテゴリーに分類されます。例えば、インテル® ディープラーニング・ブースト (インテル® DL ブースト) 搭載のインテル® Xeon® CPU、Neural Engine 搭載の Apple* CPU、テンソルコア搭載の NVIDIA* GPU、Google* TPU、AWS Inferentia*、Habana* Gaudi*、その他多くのハードウェアが、従来のハードウェア企業、クラウド・サービス・プロバイダー、AI スタートアップの組み合わせによって開発されており、現在の市場ではこれら 3 つのカテゴリーの事例がいくつか見られます。

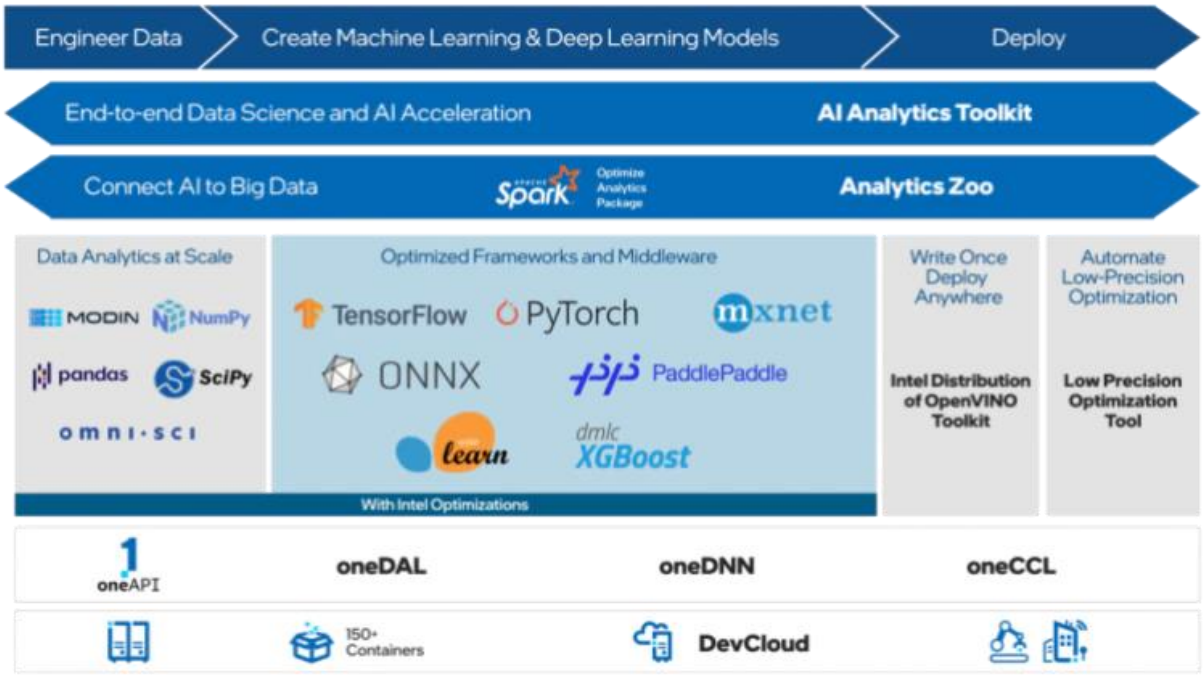
AI ハードウェアは驚異的な進歩を続けていますが、AI モデルの複雑さはハードウェアの進化を上回る速度で進んでいます。3 年ほど前、ELMo などの自然言語 AI モデルのパラメーターは「わずか」9,400 万個でしたが、2022 年現在、最大規模のモデルでは 1 兆個を超えています。AI の指数関数的な成長は、1,000 倍の計算パフォーマンスであっても、より複雑で興味深いユースケースを解決するため容易に消費できることを意味しています。世界の問題を解決し、「どこでも AI」を実現するには、ソフトウェア AI アクセラレーターによる桁違いのパフォーマンス向上が不可欠です。

ハードウェア・アクセラレーションは、自転車を最新かつ最高の機能にアップデートするようなものですが、ソフトウェア・アクセラレーションでは、超音速ジェット機のようなまったく新しい移動手段を手に入れるようなものです。

この記事では、インテル® Xeon® プロセッサー上でのソフトウェア AI アクセラレーターのパフォーマンス・データを示していますが、AI アクセラレーション搭載の CPU や GPU、専用ハードウェア AI アクセラレーターなど、ほかの AI プラットフォームでも同程度のパフォーマンス向上が達成できるでしょう。今後の記事でほかのプラットフォームでのパフォーマンス・データを紹介する予定ですが、ベンダーの皆さんもぜひソフトウェア・アクセラレーションの結果を共有してください。

AI ソフトウェア・エコシステム — パフォーマンス、生産性、オープン

AI のユースケースとワークロードがビジョン、スピーチ、推薦システムなど多岐にわたって成長し多様化し続ける中、インテルの目標は、すべての開発者、データ・サイエンティスト、研究者、データエンジニアに、エッジからクラウドまで AI を高速化する可能な限りシームレスな比類ない AI 開発および導入エコシステムを提供することです。



インテルは、オープンな標準ベースの相互運用可能なプログラミング・モデルを基盤に構築されたエンドツーエンドの AI ソフトウェア・エコシステム (英語) が、AI とデータ・サイエンス・プロジェクトを実運用にスケーリングする (英語) 鍵であると考えています。この基本的な考えは、インテルの 3 つの AI 戦略の基盤となっています。

1. **幅広い AI ソフトウェア・エコシステム上に構築する** — まず、現在の AI ソフトウェア・エコシステムを取り入れることが重要です。TensorFlow* や PyTorch*、sciKit-learn* や XGBoost、Ray* や Spark* など、ディープラーニング、マシンラーニング、データ・アナリティクスで使用されるソフトウェアを利用できるようにします。インテルは、これらのフレームワークやライブラリーを大幅に最適化し、ドロップインで 10 ~ 100 倍のソフトウェア AI アクセラレーションを実現するように設計されたインテル® プラットフォーム上で桁違いのパフォーマンス向上を達成できるように支援します。
2. **エンドツーエンドのデータサイエンスと AI ワークフローの実装** — 次に、データの準備、トレーニング、推論、展開、スケーリングなど、すべての AI ニーズに対応する各種最適化ツール (英語) を革新して提供します。例えば、エンドツーエンドのデータサイエンスとマシンラーニング・パイプラインを加速するインテル® oneAPI AI アナリティクス・ツールキット (英語)、デバイスからクラウドまでハイパフォーマンスな推論アプリケーションを展開するインテル® ディストリビューションの OpenVINO™ ツールキット、AI モデルを数千ノードのビッグデータ・クラスターにシームレスにスケーリングして分散トレーニングや推論を行う Analytics Zoo (英語) などです。
3. **比類ない生産性とパフォーマンスを実現** — 最後に、多様な AI ハードウェアに展開するため、オープンで標準ベースの統一された oneAPI プログラミング・モデルと構成ライブラリーを基盤として構築されたツールを提供します。現在、市場には多数のハードウェア AI アーキテクチャーがあり、それぞれが個別のソフトウェア・スタックを備えているため、開発者のエコシステムでは非効率的で拡張性のないアプローチとなっています。oneAPI 業界イニシアチブ (英語) は、oneAPI 仕様に対する業界全体の協力を促し、すべてのアクセラレーター・アーキテクチャーに共通の開発者エクスペリエンスを提供します。

ディープラーニング、マシンラーニング、グラフ・アナリティクスのソフトウェア AI アクセラレーター

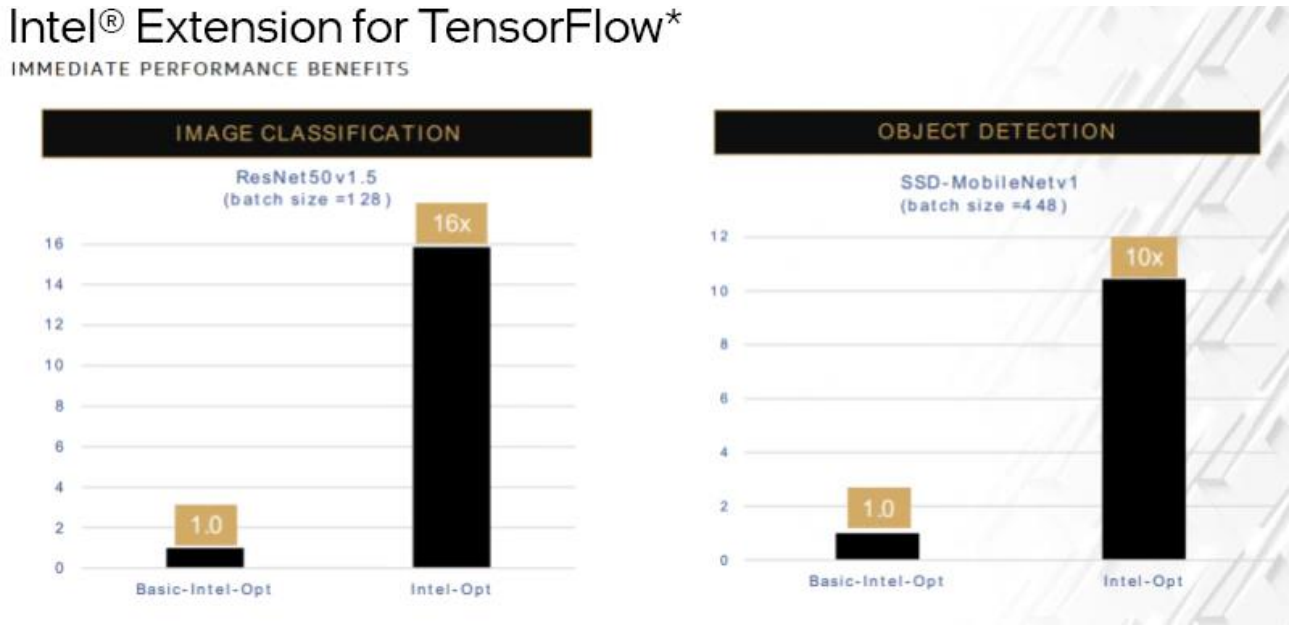
3 つの AI 戦略の最初の基盤である「ソフトウェア AI アクセラレーター」について、詳しく見ていきましょう。インテルの広範なソフトウェア最適化作業は、データ・サイエンティストが効率良くアルゴリズムを実装するシンプルな方法を提供します。インテルのライブラリーとツールは、個々の操作に対するカーネル最適化 (例えば、畳み込み操作を実装する際の SIMD レジスターの効果的な使用、ベクトル化、およびキャッシュに適したデータアクセス) と操作全体に対するグラフレベルの最適化 (バッチ正規化の畳み込み、畳み込み/ReLU 融合、畳み込み/和融合などの手法) の両方を提供します。ソフトウェア最適化手法の詳細は、こちら (英語) を参照してください。

実装の詳細が気になる方もいるかもしれませんが、インテルはこれらの最適化を抽象化し、開発者が複雑な問題に対処しなくても済むようにしました。ディープラーニング、マシンラーニング、グラフ・アナリティクスなどのインテルの最適化により、大幅なパフォーマンス向上が期待できます。

ディープラーニング

インテルのソフトウェア最適化は、インテル® oneAPI ディープ・ニューラル・ネットワーク (インテル® oneDNN) ライブラリー (英語) を使用することで、いくつかの一般的なディープラーニング・フレームワーク (英語) おいて桁違いのパフォーマンス向上をもたらします。そのほとんどは、すでにフレームワークのデフォルト・ディストリビューションにアップストリームされています。しかし、TensorFlow* と PyTorch* では、まだアップストリームされていない高度な最適化があるため、個別のインテル拡張も維持しています。

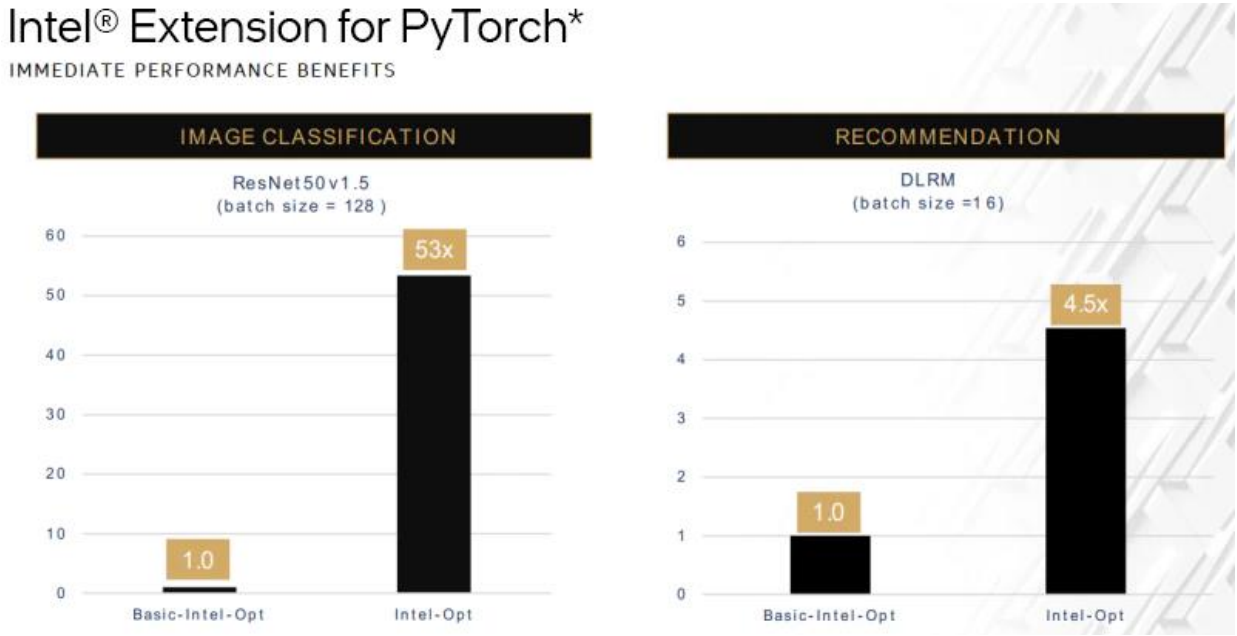
- TensorFlow* — インテルの最適化により、画像分類の推論で 16 倍、物体検出で 10 倍のパフォーマンス向上が得られました。ベースラインは、TensorFlow* Eigen ライブラリーの関数にアップストリームされている基本的なインテルの最適化を含む標準の TensorFlow* です。



システム構成: 1 ノード、2x インテル® Xeon® Platinum 8380 プロセッサー、合計 DDR メモリー 1TB (16 スロット/64GB/3200)、マイクロコード 0xd000280、インテル® ハイパースレッディング・テクノロジー有効、インテル® ターボ・ブースト・テクノロジー有効、Ubuntu* 20.04.1 LTS、5.4.0-73-generic1、インテル® SSD 900GB (OS ドライブ)、ResNet50v1.5、FP32/INT8、BS=128、https://github.com/IntelAI/models/blob/master/benchmarks/image_recognition/tensorflow/resnet50v1_5/README.md 。SSDMobileNetv1 、FP32/INT8 、BS=448 、https://github.com/IntelAI/models/blob/master/benchmarks/object_detection/tensorflow/ssd-mobilenet/README.md 。ソフトウェア: Tensorflow* 2.4.0 (FP32) & インテルの Tensorflow (icx ベース) (FP32 と INT8)。2021 年 5 月 12 日現在のインテル社内のテスト結果。

結果は異なることがあります。詳細は、http://www.intel.com/PerformanceIndex/ (英語) を参照してください。

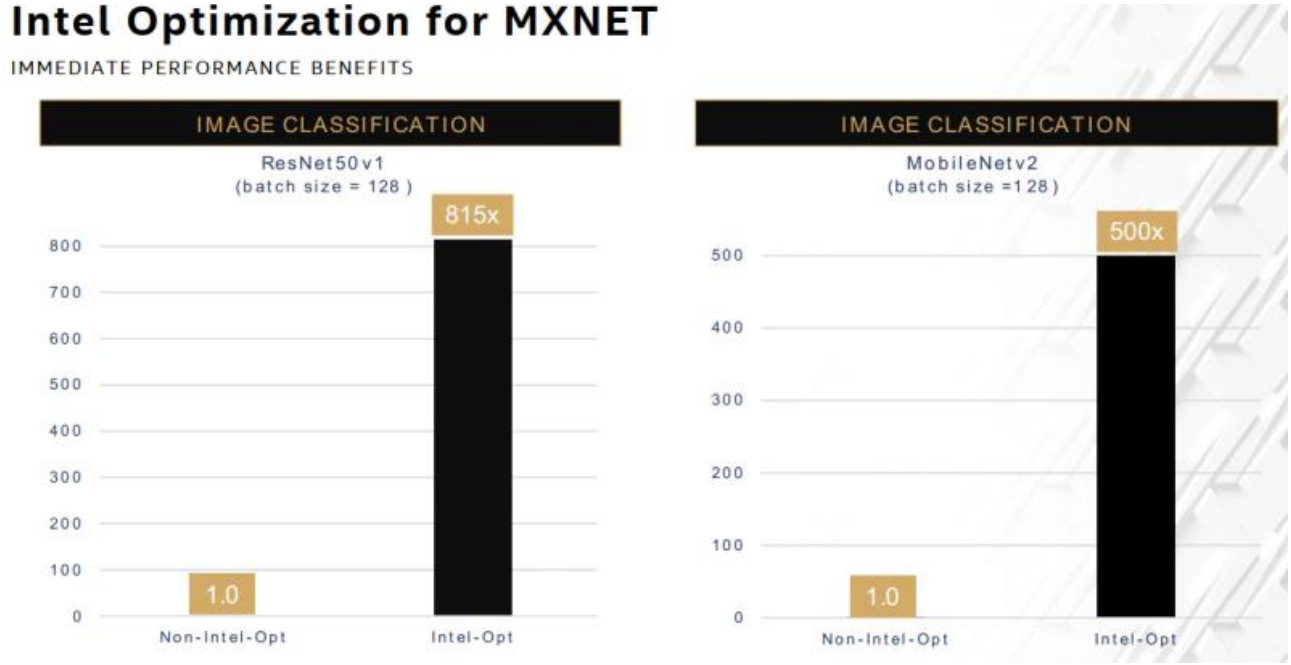
- PyTorch* — インテルの最適化により、画像分類で 53 倍、推薦システムで 5 倍近いパフォーマンス向上が得られました。oneDNN に適用されたほとんどの最適化は PyTorch* にアップストリームされていますが、まだアップストリームされていない高度な最適化があるため、個別の PyTorch* 向けのインテルの最適化も維持しています。今回の比較では、インテル® oneDNN を使用せずに、基本的なインテルの最適化だけを適用した PyTorch* で、新しいベースラインを作成しました。



システム構成: 1 ノード、2x インテル® Xeon® Platinum 8380 プロセッサー、合計 DDR メモリー 1TB (16 スロット/64GB/3200)、マイクロコード 0xd000280、インテル® ハイパースレッディング・テクノロジー有効、インテル® ターボ・ブースト・テクノロジー有効、Ubuntu* 20.04.1 LTS、5.4.0-73-generic1、インテル® SSD 900GB (OS ドライブ)、ResNet50 v1.5、FP32/INT8、BS=128、https://github.com/IntelAI/models/blob/icx-launch-public/quickstart/ipex-bkc/resnet50-icx/inference 。DLRM、FP32/INT8、BS=16、https://github.com/IntelAI/models/blob/icx-launch-public/quickstart/ipex-bkc/dlrm-icx/inference/fp32/README.md。ソフトウェア: PyTorch* v1.5 (インテル® oneDNN ライブラリーを除く) (FP32) & PyTorch* v1.5 + IPEX (icx) (FP32 と INT8)。2021 年 5 月 12 日現在のインテル社内のテスト結果。

結果は異なることがあります。詳細は、http://www.intel.com/PerformanceIndex/ (英語) を参照してください。

- MXNet — インテルの最適化により、画像分類で 815 倍と 500 倍のパフォーマンス向上が得られました。TensorFlow* や PyTorch* と異なり、MXNet にはインテル® oneDNN に適用されたすべての最適化がアップストリームされています。そのため、インテルの最適化を含まない新しいベースラインを作成して比較しました。

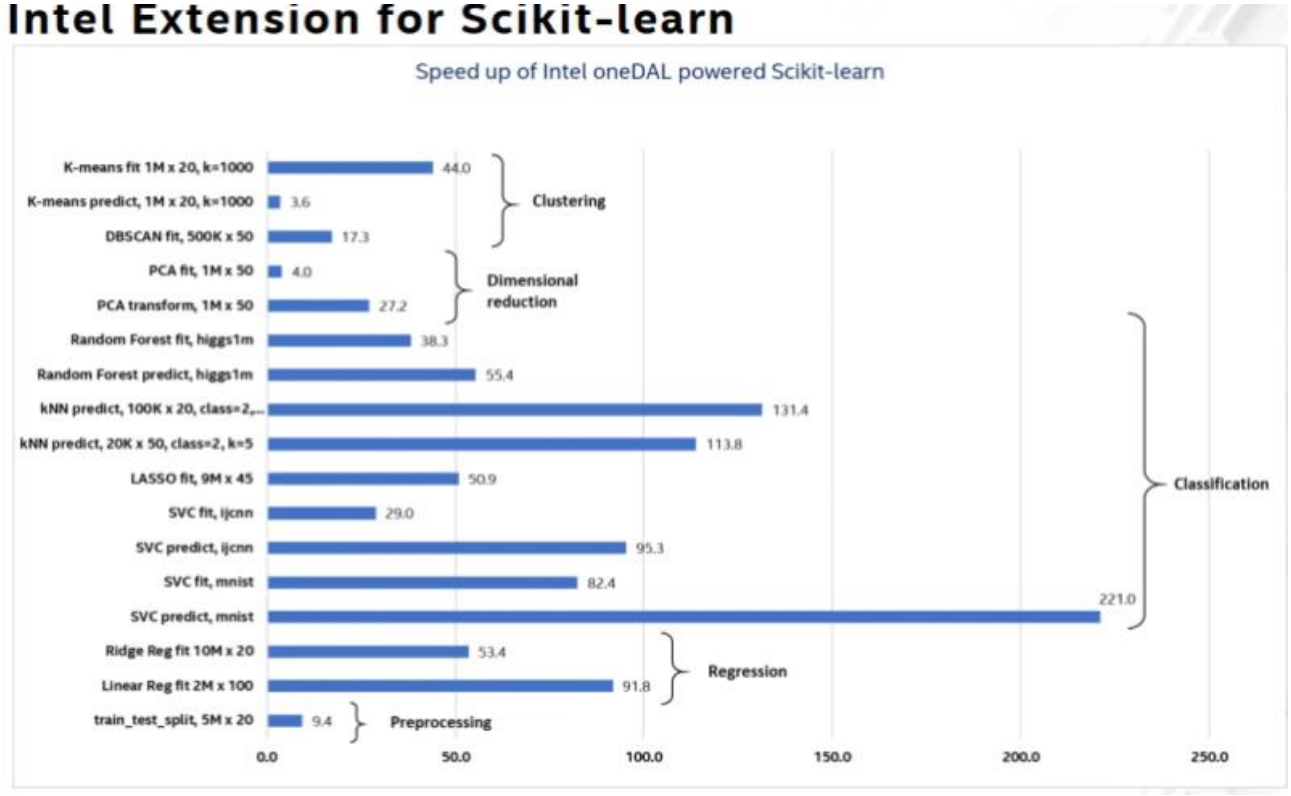


システム構成: 1 ノード、2x インテル® Xeon® Platinum 8380 プロセッサー、合計 DDR メモリー 1TB (16 スロット/64GB/3200)、マイクロコード 0xd000280、インテル® ハイパースレッディング・テクノロジー有効、インテル® ターボ・ブースト・テクノロジー有効、Ubuntu* 20.04.1 LTS、5.4.0-73-generic1、インテル® SSD 900GB (OS ドライブ)、ResNet50v1、FP32/INT8、BS=128、https://github.com/apache/incubator-mxnet/blob/v2.0.0.alpha/python/mxnet/gluon/model_zoo/vision/resnet.py。MobileNetv2、FP32/INT8、BS=128、https://github.com/apache/incubator-mxnet/blob/v2.0.0.alpha/python/mxnet/gluon/model_zoo/vision/mobilenet.py。ソフトウェア: MXNet 2.0.0.alpha (インテル® oneDNN ライブラリーを除く) (FP32) & MXNet 2.0.0.alpha (FP32 と INT8)。2021 年 5 月 12 日現在のインテル社内のテスト結果。

結果は異なることがあります。詳細は、http://www.intel.com/PerformanceIndex/ (英語) を参照してください。

マシンラーニング

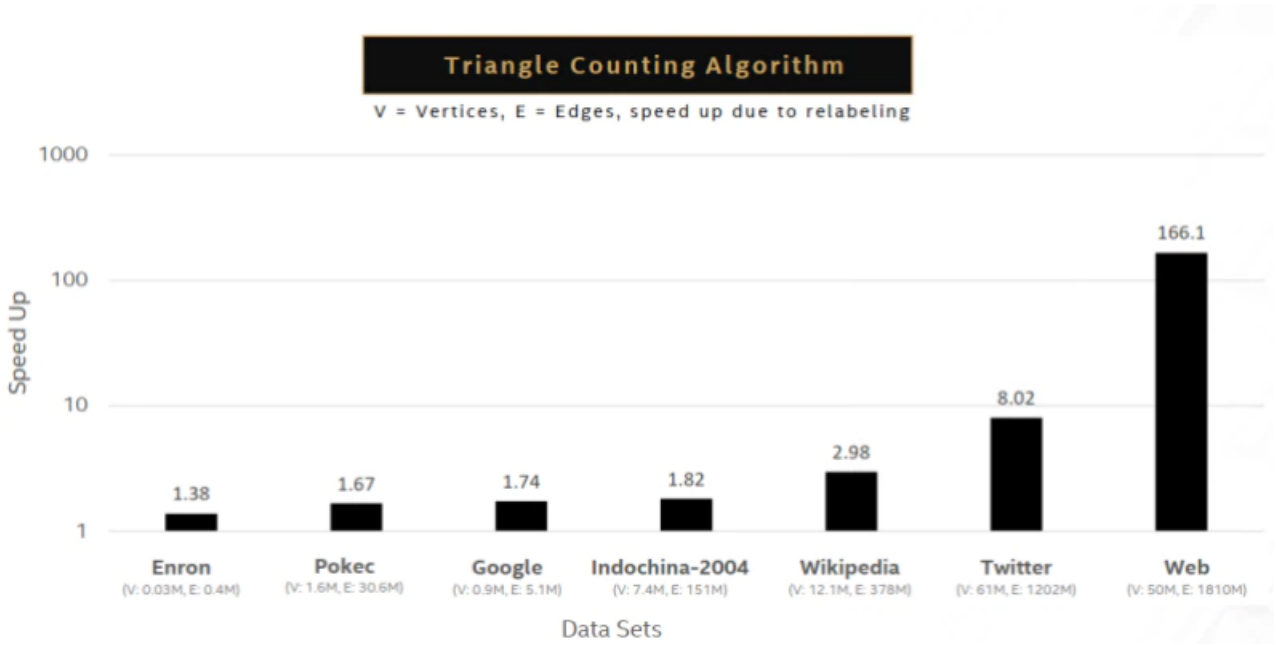
scikit-learn* は、Python* 用の人気の高いマシンラーニング・ソフトウェア・ライブラリーです。サポート・ベクトル・マシン、ランダムフォレスト、勾配ブースティング、K 平均法など、さまざまな分類、回帰、クラスタリングのアルゴリズムを提供します。これらの一般的なアルゴリズムのパフォーマンスを、最大 100 ~ 200 倍向上 (英語) させることができました。これらのパフォーマンス向上は、scikit-learn* 向けインテル® エクステンションとインテル® oneAPI データ・アナリティクス・ライブラリー (インテル® oneDAL) を使用することで達成できます。



システム構成: インテル® Xeon® Platinum 8276L プロセッサー @ 2.20GHz、2 ソケット、ソケットごとに 28 コア。詳細は、http://www.intel.com/PerformanceIndex/ (英語) を参照してください。詳細: https://medium.com/intel-analytics-software/accelerate-your-scikit-learn-applications-a06cacf44912 (英語)、https://medium.com/intel-analytics-software/save-time-and-money-with-intel-extension-for-scikit-learn-33627425ae4 (英語)、および https://medium.com/intel-analytics-software/leverage-intel-optimizations-in-scikit-learn-f562cb9d5544 (英語)。

グラフ・アナリティクス

グラフ・アナリティクスとは、ソーシャル・ネットワーク、インターネットと検索、Twitter、Wikipedia などの大規模なグラフ・データベースにおいて、項目間の関係の強さと方向の調査に使用されるアルゴリズムを指します。広く使われているグラフ・アナリティクス・アルゴリズムとして、単一ソース最短パス、幅優先探索、連結成分、ページランク、媒介中心性、三角形カウントなどがあります。例えば、インテルの最適化により、三角形カウント・アルゴリズムは大幅にパフォーマンスが向上します。グラフが大きくなるほど最適化の効果も大きくなり、5,000 万頂点、18 億エッジに迫る最大規模のグラフでは 166 倍のパフォーマンス向上が見られます。こちらの記事 (英語) では、ほかのいくつかのグラフ・アナリティクス・アルゴリズムに対するインテルの最適化について、より詳しく説明しています。



システム構成: インテル® Xeon® Platinum 8280 プロセッサー @ 2.70GHz、2 ソケット、ソケットごとに 28 コア、インテル® ハイパースレッディング・テクノロジー有効。詳細は、http://www.intel.com/PerformanceIndex/ (英語) を参照してください。データセット: https://gihub.com/sbeamer/gapbs | https://snap.Stanford.edu/data。

どこでも AI — ソフトウェア AI アクセラレーションの応用

AI で問題を解決するには、エンドツーエンドのワークフローが必要です。データから開始し、それぞれのユースケースには独自の AI データ・パイプラインがあります。AI 実務者は、データを取り込み、必要に応じてマシンラーニング (ML) を使って特徴エンジニアリングで前処理し、DL や ML を使ってモデルをトレーニングして展開します。インテル® oneAPI AI アナリティクス・ツールキット (英語) は、これらのパイプラインのすべてのフェーズを加速し、ソフトウェア AI アクセラレーションによって大幅なスピードアップを達成するハイパフォーマンス API と Python* パッケージを提供します。こちらの記事 (英語) では、インテル® oneAPI AI アナリティクス・ツールキットにより、データ・サイエンティストが AI パイプラインの高速化を実現した 2 つの実例 (米国国勢調査と PLAsTiCC 天文分類) を詳しく説明しています。

ソフトウェア AI アクセラレーターは、すでに AI の成長と各種ドメインやユースケースへの普及に不可欠なパフォーマンス向上を実現していることが分かりました。今後さらに多くのことを実現できるでしょう。インテルは、さらにソフトウェア AI アクセラレーションを推進するため、コンパイラー・テクノロジー、メモリー最適化、分散コンピューティングに取り組んでいます。AI ソフトウェア・コミュニティー全体が協力し、インテルや他のハードウェア・ベンダーが低レベルのソフトウェアやフレームワークの最適化を率先し、ソフトウェア・ベンダーが高レベルの最適化を主導し、それらを業界標準の中間表現に統合することで、ソフトウェア AI アクセラレーターの真価を引き出すことができるでしょう。

AI システム構築者はソフトウェアをより重視し、開発者と実務者は AI パフォーマンス高速化の可能性をさらに追求することをお勧めします。

(1) ディープラーニングやマシンラーニングのフレームワーク (TensorFlow*、PyTorch*、MXNet、XGBoost、scikit-learn* など) を使用する際は、常にすでに多くのインテルの最適化がアップストリームされている最新版を使用します。

(2) さらに高いパフォーマンスを達成するには、最新の最適化をすべて含み、既存のワークフローと完全に互換性のあるフレームワーク (英語) のインテル拡張を使用します。

ドロップインのフレームワーク最適化およびパフォーマンス最適化を利用可能なインテル® AI ソフトウェア (英語) のエンドツーエンドのツールの詳細を確認し、AI ワークフローのパフォーマンスを最大 100 倍向上できることを実感してください。

ソフトウェア AI アクセラレーターとハードウェア AI アクセラレーターを組み合わせて「どこでも AI」の未来を実現することで、よりスマートな、コネクティビティーに優れた、すべての人にとってより良い世界に到達できるでしょう。

法務上の注意書き

性能は、使用状況、構成、その他の要因によって異なります。詳細については、http://www.intel.com/PerformanceIndex/ (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。 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。実際の費用と結果は異なる場合があります。インテルのテクノロジーを使用するには、対応したハードウェア、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。

** これは、(a) Netflix*、Amazon Prime*、Disney* などの主要なストリーミング・プロバイダーのストリーミング・コンテンツの時間および運営国、(b) AWS*、Microsoft* Azure*、Google Cloud* などの米国の主要 CSP におけるコンピューター・ビジョンおよび NLP AI サービスの使用コストに関する公開情報に基づく近似値です (ただし、これに限定されるものではありません)。この推定値は、問題の範囲やコスト削減の可能性を説明するためのものであり、インテルはその正確性を保証するものではありません。実際の費用と結果は異なる場合があります。

関連記事

Databricks Runtime ML のスピードアップ
読む (英語)

1 行のコード変更で pandas、scikit-learn*、TensorFlow* のパフォーマンスを向上
読む (英語)

関連ビデオ

AI アナリティクス・パート 1: エンドツーエンドのデータサイエンスとマシンラーニングの高速化を最適化
視聴する

AI アナリティクス・パート 2: 第 3 世代インテル® Xeon® スケーラブル・プロセッサー上でディープラーニング・ワークロードを強化
視聴する

AI アナリティクス・パート 3: エンドツーエンドのマシンラーニング・ワークフローの最適化ステップ
視聴する

ソフトウェアを入手

インテル® oneAPI AI アナリティクス・ツールキット (英語)

最適化された DL フレームワークとハイパフォーマンスの Python* ライブラリーでエンドツーエンドのマシンラーニングとデータ・サイエンス・パイプラインを高速化

今すぐ入手 (英語)
すべてのツールを見る (英語)