Windows* ML: インテル® ハードウェア上での AI の高速化

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「Windows* Machine Learning: AI Acceleration on Intel® Hardware」の日本語参考訳です。

はじめに

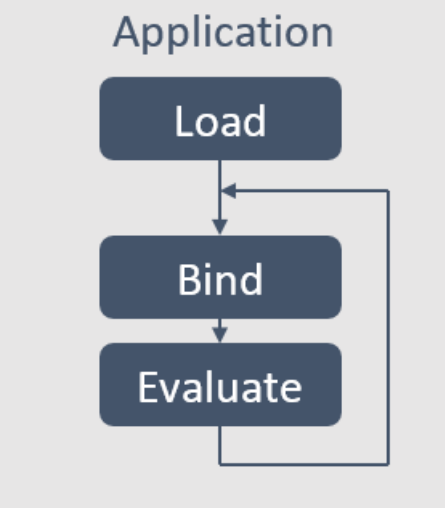
人口知能 (AI) は、サーバー、デスクトップ、エッジ市場に拡大しており、インテルはこれらの市場において AI ソリューションを展開しています。インテル® Xeon® プロセッサー、インテル® Core™ プロセッサー、Intel Atom® プロセッサーなどの AI 処理の基礎を提供する CPU インフラストラクチャーから、インテル® Iris® グラフィックスによる専用のアクセラレーション、低消費電力のインテル® Movidius™ ビジョン・プロセシング・ユニット (VPU)、新しいインテルのガウス・ネットワーク・アクセラレーター (GNA)、Mobileye* 車載機器テクノロジー、インテル® FPGA カスタム統合まで、インテルの AI 製品は多種多様なアプリケーションにわたっています。

Windows* ML (Machine Learning) は、エッジ上の Windows* で動作する推論エンジンです。開発インターフェイスは非常にシンプルで、インテル® ハードウェア向けに最適化されています。

インテルは、Windows* ML を使用したハードウェアの最適化が、モデルの評価において最先端のアクセラレーションを達成できるように Microsoft* と緊密に協力しています。

Windows* ML: 開発者にとっての利点

Windows® 10 の 2018 年 4 月のアップデートで利用可能な Windows* ML API には、非常に単純なプログラミング・モデルが含まれています。アプリケーションは、訓練済みのモデルをロードして、データをモデルにバインドし、データに対してモデルを評価するだけで済みます。その他のアクセラレーションは、インテル® ハードウェア上で最高のパフォーマンスを達成できるようにレイヤーで最適化されています。

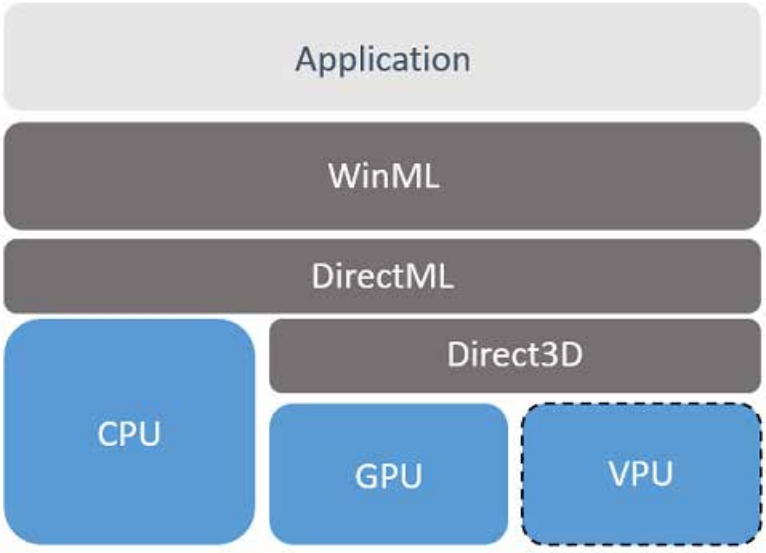


インテル® ハードウェア上での Windows* ML

インテル® ハードウェア上では、Windows* ML スタック (以下の図) はインテル® アドバンスト・ベクトル・エクステンション 512 (インテル® AVX-512) CPU 命令セットと DirectX* 12 計算パイプラインの両方に影響を与え、統合グラフィックス上での実行を高速化します。

このスタックには、市場のほかの推論エンジンと同様に、モデルを理解するための Windows* ML 推論エンジンがあります。Direct ML 抽象化レイヤーは、評価するターゲット・ハードウェアを選択して、手動で最適化されたインテル® AVX-512 命令セットベースのモデルの CPU 実装を実行するか、Direct3D* DirectX* 12 インターフェイスを介して高水準のシェーディング言語を使用するシェーダーをハードウェアへ発行します。

インテルは、新しいインテル® Movidius™ VPU により専用の低消費電力の AI アクセラレーターを提供します。また、Windows® 10 の Fall OS アップデートにおいて、DirectX* 12 レイヤーの新しい Meta Command インターフェイスを使用してモデルレベルでオペレーターを高速化するため、インテルは Microsoft* と共同で取り組んでいます。

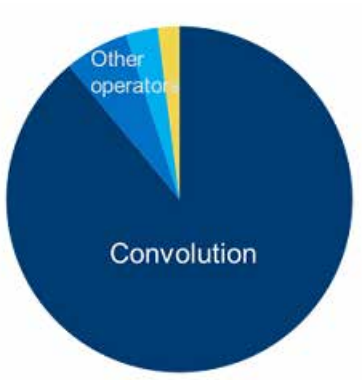


オペレーターごとの高速化 (Meta Command) の使用

インテル® インテグレーテッド・グラフィックス・デバイス上で最高のパフォーマンスを達成するため、インテルは Microsoft* と緊密に協力しています。モデルが評価されると、Windows* ML ソフトウェア・スタックは、特定のモデル操作の最適化されたバージョンが利用できるかどうかをドライバーに問い合わせます。利用可能な場合、高速化された操作を使用します。アプリケーションの変更は不要です。

これらの操作はドライバーの内部にあり、次の OS アップデートとドライバー・アップデートでアプリケーションに提供されます。これらの操作は、実行ユニットでの効率を高めてインテル® インテグレーテッド・グラフィックスを最適に使用するように手動で最適化されたカーネルと、ハードウェア向けの最適なキャッシュ手法により実行されます。これらのカーネルは、インテルのエキスパートによってインテル® ハードウェア向けにチューニングされており、高速化された新しい機能を制御するため開発者がアプリケーションを変更する必要はありません。

実際の AI イメージング・モデルに基づいて、最初に高速化された操作は畳み込みです。このオペレーターは、基本的にイメージ全体を処理する大規模な行列乗算で、非常に高価です。初期結果では十分に効果が得られ、時間の経過とともにさらに向上することが期待されます。



より多くのオペレーターが高速化され、簡単な OS とドライバー・レベルのアップデートにより改善されたエコシステム・モデルがさらに高速化されることで、全体的な結果も時間とともに向上します。

そのため、インテル® インテグレーテッド・グラフィックス上で最高のパフォーマンスを達成するには、新しい Windows* ML API を使用し、今後提供される Meta Command でさらなる最適化を利用して、アプリケーションを変更することなくパフォーマンスを向上できます。

今後も、インテル® グラフィックス上での Windows* ML に関する最新情報をお見逃しなく。

コンパイラーの最適化に関する詳細は、最適化に関する注意事項を参照してください。