OpenVINO™ を使用してクラウドのトレーニング済みモデルのパフォーマンスを最大限に引き出す

この記事は、Medium に公開されている「Have you trained a model in the cloud? Use OpenVINO™ to maximize performance!」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。

1. はじめに

この数年で、ディープラーニングは、画像認識や音声認識、自然言語処理など、幅広いアプリケーションにとって強力なツールとなりました。ディープラーニング・モデルの機能が進歩するにつれて、カメラ、ドローン、ロボットなどのエッジデバイスにデプロイするため、これらのモデルを最適化する必要性が高まっています。

ディープラーニング・モデルのパフォーマンスを向上させて、異なるハードウェア・プラットフォームで効率良く動作させるには、OpenVINO™ などのツールを使用してデプロイのためにモデルを最適化することが重要です。デプロイのためにディープラーニング・モデルを最適化すると、ターゲット・ハードウェア・プラットフォームの計算リソースが効率良く使用され、最適なパフォーマンスが得られます。この最適化は、計算リソースが限られていて、モデルのパフォーマンスが重要なエッジデバイスにディープラーニング・モデルをデプロイする場合に特に重要です。

ディープラーニング・モデルをトレーニングするさまざまなサービスを提供するクラウドベースのプラットフォームはいくつかあります。これらのサービスには通常、モデルをトレーニングする計算能力、データとモデルを保存するストレージサービス、マシンラーニング・モデルを構築、トレーニング、デプロイするエンドツーエンドのプラットフォームへのアクセスが含まれます。プラットフォームには、Amazon SageMaker*、Google Cloud* AI Platform、Microsoft* Azure* Machine Learning、IBM Watson* Studio などが含まれます。

クラウド・プラットフォームを選択してモデルをトレーニングしたら、OpenVINO™ を使用してモデルを最適化し、さまざまなハードウェア・プラットフォームにデプロイできます。ここでは、OpenVINO™ を使用して、トレーニング済みのディープラーニング・モデルの推論のパフォーマンスを最適化する方法の概要を説明します。この

記事を最後までお読みになることで、OpenVINO™ を使用してディープラーニング・モデルを最適化およびデプロイする方法を理解できるようになります。

2. OpenVINO™ はさまざまな AI フレームワークとハードウェアをサポート

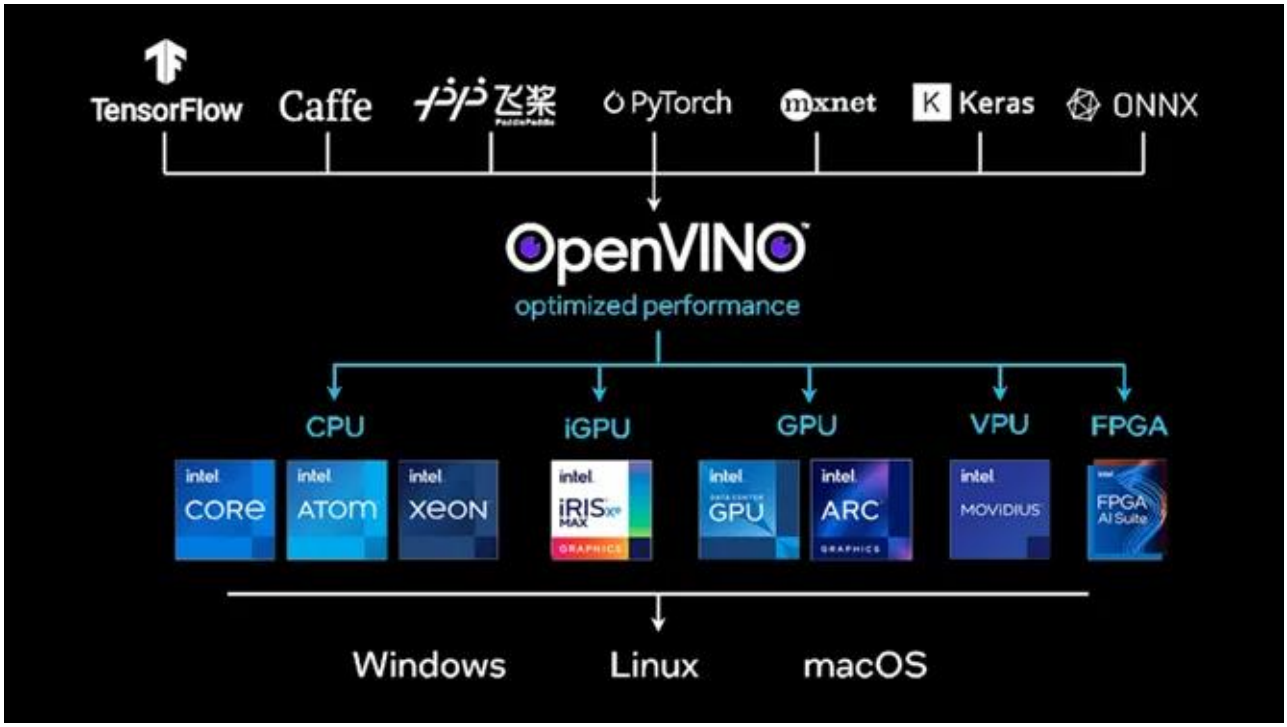

図 1: OpenVINO™ はさまざまなディープラーニング・フレームワークとハードウェアをサポート

図 1 に示すように、OpenVINO™ は、開発者がインテルのハードウェアでディープラーニング・アルゴリズムを効率良く実行できるように支援するツールキットです。ディープラーニングの推論アプリケーションで低いレイテンシーと高いスループットを実現する、ディープラーニングの推論オプティマイザーとランタイムが含まれています。OpenVINO™ は、TensorFlow*、PyTorch*、ONNX*、MxNET、Caffe* など、複数のディープラーニング・フレームワークをサポートしています。

OpenVINO™ は、CPU、GPU、VPU、FPGA を含む、さまざまなインテルのハードウェアでディープラーニングの推論を実行するように最適化されています。OpenVINO™ を使用することにより、開発者は、これらのインテルのハードウェア・プラットフォームのパフォーマンスと電力効率を活用して、ディープラーニング・モデルを高速に効率良く実行できます。

OpenVINO™ の「1 つのコードでどこでもデプロイできる」理念により、開発者は、OpenVINO™ を使用してディープラーニングの推論コードを一度記述すれば、コードを変更することなく、異なるインテルのハードウェア・プラットフォームにデプロイできます。開発者は、プラットフォームごとにカスタムコードを記述することなく、特定のハードウェア・プラットフォーム向けにモデルを簡単に最適化できます。そして、低電力 CPU を搭載したエッジデバイスから強力な GPU を搭載したハイパフォーマンス・サーバーまで、広範なデバイスにディープラーニング・モデルを簡単にデプロイできます。

3. OpenVINO™ モデルの最適化とデプロイ

OpenVINO™ を使用して特定のハードウェア・プラットフォーム向けにディープラーニング・モデルを最適化する手順を説明します。最初に、OpenVINO™ モデル・オプティマイザーを使用して、OpenVINO™ がディープラーニング・モデルを表現するために使用する一般的な形式である、中間表現 (IR) 形式にモデルを変換します。

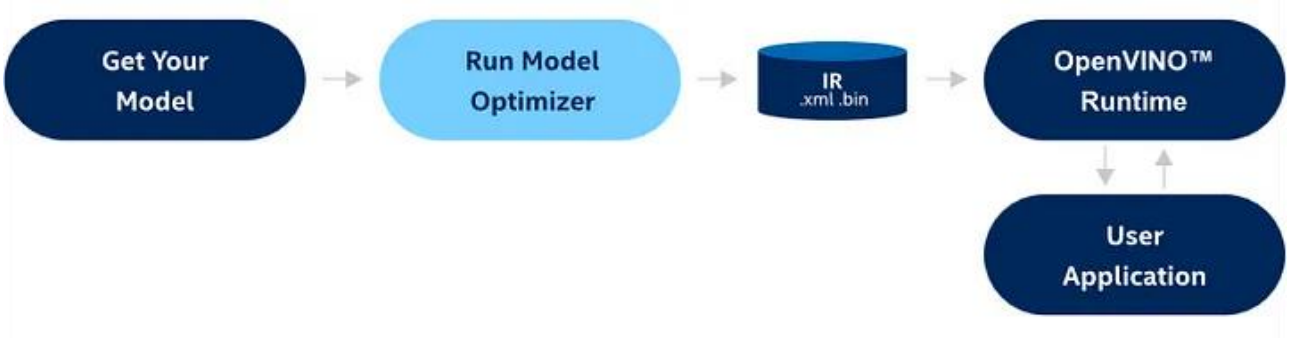

図 2: OpenVINO™ モデルの変換と推論のワークフロー

IR 形式にモデルを変換すると、OpenVINO™ 推論エンジンを使用して特定のハードウェア・プラットフォーム向けに最適化できます。このプロセスには、デバイスで利用可能な CPU コア数やメモリー量など、ターゲット・ハードウェアの特性に応じたモデルのチューニングが含まれます。

最後に、最適化したモデルを目的のハードウェア・プラットフォームにデプロイします。具体的には、OpenVINO™ 推論エンジンを使用してモデルをユーザー・アプリケーションと統合します。これで、アプリケーションは、モデルを使用して入力データにディープラーニングの推論を実行し、ターゲット・ハードウェア・プラットフォームのパフォーマンスと電力効率を活用することができます。あるいは、OpenVINO™ モデルサーバーを使用して推論リクエストを処理することもできます。

例として、ユーザーが Amazon Sagemaker* を使用して AWS* クラウドで TensorFlow* トレーニング・パイプラインを実行しているシナリオを検討しました。目的のモデルのトレーニング目標を達成し、モデルをデプロイする準備ができたら、OpenVINO™ を使用して、エッジデバイス、クラウド・インスタンス、IoT デバイスなどのインテルのデバイス、またはその他のインテルのプラットフォームで、最良の推論パフォーマンス (スループット/レイテンシー) が得られるようにトレーニング済みモデルを最適化できます。OpenVINO™ を使用するプロセスは、Azure*、GCP、その他のクラウド・サービス・プロバイダーでも同じです。

図 3 の左側は、トレーニング環境とパイプライン、およびトレーニング済みモデルを最適化して AWS* S3 バケットにプッシュする OpenVINO™ モデル・オプティマイザーのタスクを示しています。右側は、推論出力を生成するインテルのデバイスの、Ubuntu* ベースの Docker* コンテナ内の OpenVINO™ モデル推論ワークフローを示しています。

注: Amazon Sagemaker* で示されている OpenVINO™ モデル最適化ブロックは、ワークフローの推論側でも実行できます。

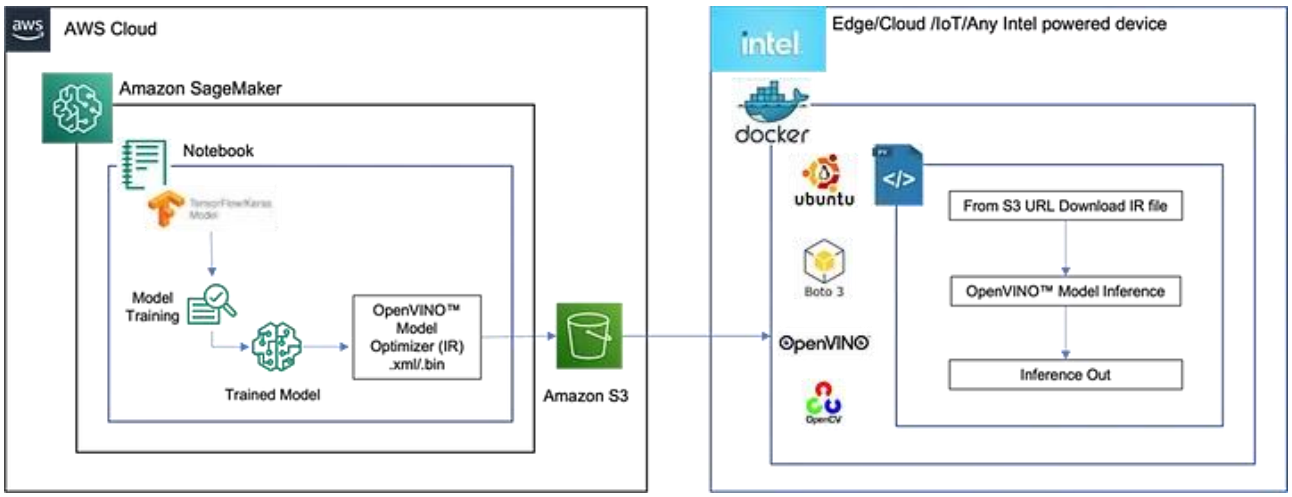

図 3: Amazon SageMaker* と OpenVINO™ を使用したデプロイシナリオの例

4. 関連情報

1. インテル® ディストリビューションの OpenVINO™ ツールキット
2. OpenVINO™ ツールキット入門 (英語)
3. Jupyter* Notebook のコレクション (英語)
4. OpenVINO™ サンプルコード (英語)
5. OpenVINO™ モデル・オプティマイザー (英語)
6. OpenVINO™ ツールキット API (英語)
7. OpenVINO™ Model Zoo (英語)
8. AWS* クラウド/Amazon Sagemaker* の OpenVINO™ Amazon* Machine Image (AMI) (英語)
9. OpenVINO™ モデルサーバー (英語)

OpenVINO™ ツールキットとは

AI を加速する無償のツールである OpenVINO™ ツールキットは、インテルが無償で提供しているインテル製の CPU や GPU、VPU、FPGA などのパフォーマンスを最大限に活用して、コンピューター・ビジョン、画像関係をはじめ、自然言語処理や音声処理など、幅広いディープラーニング・モデルで推論を最適化し高速化する推論エンジン/ツールスイートです。

OpenVINO™ ツールキット・ページでは、ツールの概要、利用方法、導入事例、トレーニング、ツール・ダウンロードまでさまざまな情報を提供しています。ぜひ特設サイトにアクセスしてみてください。

https://www.intel.co.jp/content/www/jp/ja/internet-of-things/openvino-toolkit.html