

ホワイトペーパー

OpenVINO™ ツールキット
インテル® DevCloud



ディープラーニングを使って X線画像から疾患を推定するシステムを構築 最適化や量子化で100倍の性能向上を確認

ディープラーニングを用いて胸部X線画像から疾患を推定するシステムを対象に、
最適化と量子化を適用した性能評価を実施しました。インテル® DevCloud を開発および実行環境とし、
開発ツールには OpenVINO™ ツールキットを利用しました。

国立研究開発法人理化学研究所
光量子工学研究センター
データサイエンティスト

種石 慶 氏

インテル株式会社
アジア・パシフィック・ジャパン
データセンター・グループ・セールス
AIテクニカルソリューションスペシャリスト

大内山 浩

背景

ディープラーニングの応用が医療分野にも広がっています。代表的なのが、眼底画像、X線画像、CT 画像、MRI 画像（以下、医用画像）などの読影です。これら画像に畳み込みネットワーク（CNN）を用いた画像認識（画像分類）を適用し、疾患名や病変部位を推定するシステムの開発および実用化が進められています。

読影の最終判断はもちろん医師が担いますが、ディープラーニングを用いたシステムが参考となる情報を示すことによって、次のようなメリットが得られると考えられます。

- (1) 医師および一次読影を担当する読影医の負担軽減（特に画像枚数の多い健康診断）。
- (2) 読影医の不足などによる読影基準のばらつきの標準化および均一化の推進。
- (3) ディープラーニングならではの画像分類精度を生かした、ごくわずかな病変の検出。

目的

医用画像を対象にしたディープラーニングの応用では3つの課題が指摘されています。1つがラベル付けされた医用画像の収集で、被験者のプライバシーに配慮しながら学習に必要なだけの枚数の画像を集めなければなりません。2つ目が精度の高い推論モデル（分類や注釈の予測）の開発で、多くの時間と計算コストを必要とします。

ただし、これらの課題に対しては対応が進んでおり、後述する NIH（アメリカ国立衛生研究所）のデータセットのように、いくつかの研究機関が症例ラベルを付加した医用画像データセットの公開を始めています。その結果、誰もが推論モデルの研究開発を進められる環境が整ってきました。そして実際に、後述する CheXNet* に代表される高精度な推論モデルが開発・公開されています。

3つ目の課題が推論処理における計算コストの問題です。精度の高い推論モデルが得られたとしても、推論処理のコストが高ければ、医療現場への導入は進みません。

将来の普及のためには、医用画像の推論処理を、高性能（高額）なサーバーや AI アクセラレーターを用いることなく、実用的な計算コスト（コンピューティング・リソース）で実現することが求められます。

計算コストを抑える有力な手段の1つが推論モデルの最適化および量子化(軽量化)です。計算量を削減できれば、一般的なサーバーや、条件によってはクライアントPCでの処理(エッジ・コンピューティング)が可能になり、高性能なサーバー設備や高性能なクラウドサービスを調達しなくても済みます。

こうした課題を踏まえて、国立研究開発法人理化学研究所(理研)光量子工学研究センターの種石慶氏によるワークフローの提供の下、推論モデルの最適化および量子化の評価検証を行いました。主な手順は次のとおりです(図1)。

- (a) 公開された推論モデルの利用を前提として、米スタンフォード大学によって開発された胸部X線画像の推論モデルである「CheXNet*」¹を用いる。
- (b) ディープラーニングのツールキットである「OpenVINO™ツールキット」を用いてモデルの最適化および量子化を図る。
- (c) 開発にはディープラーニングの開発環境(サンドボックス)である「インテル® DevCloud」を利用する。
- (d) ベンチマークリングにより性能を確認する。

実装

・データセットと推論モデル

今回の評価・検証では、米スタンフォード大学のAndrew Ng氏のチームが開発した推論モデル CheXNet* を対象に、モデル最適化および量子化を行いました(一部 CheXNet* の出力層を変更した改良版を用いています)。なお、CheXNet* は、CNNの構造としては、基本的な画像処理用ニューラル・ネットワークの1つである121層のDenseNet* (DenseNet-121)を採用しています。

CheXNet* が教師あり学習データとして使用したのが、NIH Clinical Centerが2017年に公開した胸部X線画像のデータセット「ChestX-ray14」²です。ChestX-ray14は30,805人の患者から得た延べ112,120枚の胸部X線画像で構成され、肺炎、胸水、無気肺、心肥大、結節、気胸、浸潤影、肺水腫、気腫、肺線維症、ヘル

ニア、胸膜肥厚など、14の疾患がラベル付けされています(1枚の画像に複数の疾患がラベル付けされている場合もあります)。ただし、新型コロナウイルス感染症(COVID-19)が発生する前に作成されたデータセットですので、COVID-19のラベルが付いた肺炎画像は含まれていません。

・最適化および量子化のツール

CheXNet* の最適化および量子化にはインテルが無償で提供している OpenVINO™ ツールキット³を使用しました。

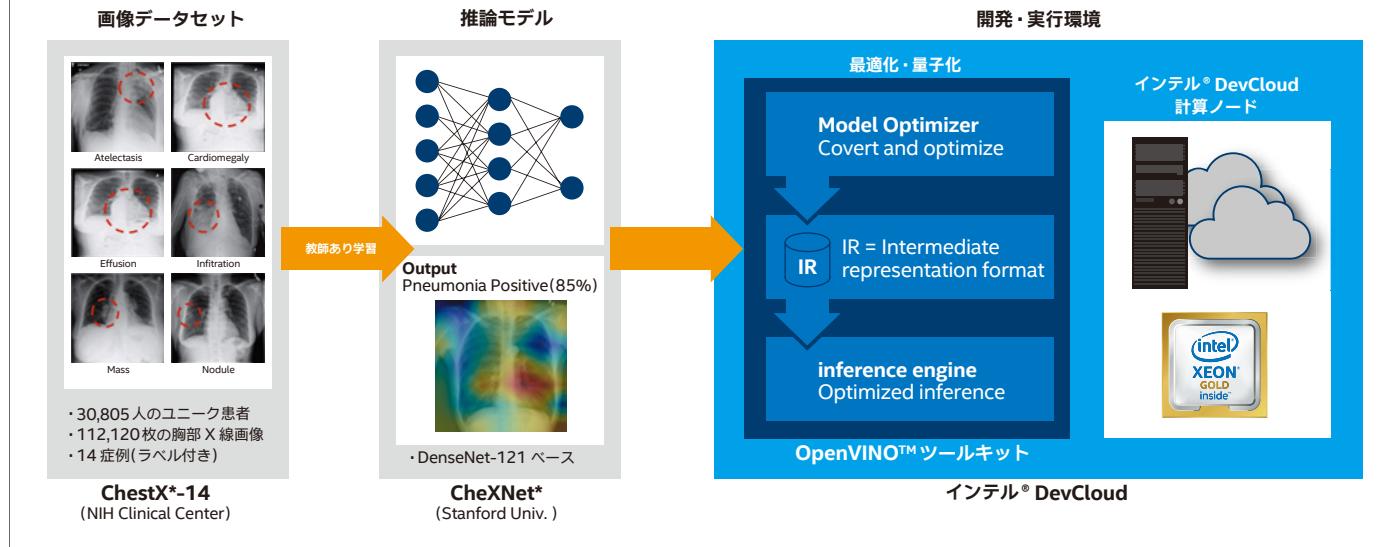
OpenVINO™ ツールキットはディープラーニングを用いたビジョン・アプリケーションの開発を支援するツールキットで、他のディープラーニング・フレームワークで開発された推論モデルを、インテル® Xeon® スケーラブル・プロセッサー、インテル® Core™ プロセッサー・ファミリー、インテル® FPGA、インテル® Movidius™ ビジョン・プロセシング・ユニットなど、さまざまなハードウェアに効率的に展開する機能を備えています。動作環境としては、Microsoft* Windows* 10、Linux*、macOS* がサポートされています(Raspbian OS* でも一部機能を除いて動作可)。

OpenVINO™ ツールキットの特徴的な機能の1つが「Model Optimizer」です。外部から ONNX 形式などを通じて読み込んだ推論モデルに対して最適化を行い、IR(中間表現)形式として出力します。IR 形式のモデルは OpenVINO™ ツールキットが提供するハードウェア・プラットフォームごとの「Inference Engine」(API セット)上で実行されます。

OpenVINO™ ツールキットのもう1つの特徴が32ビット浮動小数点(FP32)から8ビット整数(INT8)への量子化を行う「DL Workbench*」(キャリブレーション・ツール)機能です。なお、DL Workbench* はオプションとなっていて、別途インストールが必要です(Microsoft* Windows* で動かす場合はコンテナ環境が必要、macOS* には非対応)。

DL Workbench* によって INT8 に量子化されたモデルは、単

図1. 胸部X線画像を対象にしたディープラーニング推論モデルの最適化および量子化の大まかな流れ



に数値演算処理が軽量化されるだけではありません。第1世代のインテル® Xeon® スケーラブル・プロセッサー (Skylake マイクロアーキテクチャー)上では、8ビットの畳み込み積和演算は、AVX-512 命令セットに含まれる VPMADDUBSW、VPMADDWD、VPADDD 命令を使用して高速に実行されます。

さらに第2世代インテル® Xeon® スケーラブル・プロセッサー (Cascade Lake マイクロアーキテクチャー) であれば、8ビットの畳み込み積和演算は新たに追加されたベクトル・ニューラル・ネットワーク命令セット (VNNI) に含まれる VPDPBUSD 命令のみで実行されるため、さらなる性能向上が期待できます。

・開発環境および実行環境

最適化と量子化の効果を検証するには実行環境が必要です。OpenVINO™ ツールキットをインストールした開発環境をそのまま実行環境として用いる方法もありますが、今回の評価検証ではインテルが開発者向けに提供しているインテル® DevCloud⁴ を利用することにしました。

インテル® DevCloud は、インテル® Xeon® スケーラブル・プロセッサー やインテル® Core™ プロセッサー・ファミリー のほか、AI アクセラレーターであるインテル® Movidius™ ビジョン・プロセシング・ユニットなど、さまざまなハードウェアを取り揃えたクラウドサービスで、いわゆるサンドボックスに相当します。

インテル® DevCloud の各ノード (サーバー) には OpenVINO™ ツールキットが搭載されていて、IR 形式の推論モデルは指定されたノード上の Inference Engine 上で実行されます。最新のインテル® Xeon® スケーラブル・プロセッサーを含むさまざまなハードウェアで性能を検証できるのが特徴で、ユーザーはハードウェアを自前で揃える必要がありません。インテル® DevCloud の詳細については別欄をご参照ください。

今回は、VNNI が利用できる第2世代のインテル® Xeon® Gold 6258R プロセッサー (動作周波数 2.70GHz) を選択し、ベンチマークを行いました。

結果と評価

CheXNet* を対象にしたベンチマークの結果を図2に示します。まず、オープンソースの PyTorch* を使って CheXNet* をほぼそのままの形で実行した場合、ChestX-ray14 の画像あたりの分類時間は 17.79 秒でした。これに対して OpenVINO™ ツールキットの Model Optimizer で最適化を行ったところ、実行時間はわずか 1.65 秒に短縮され、最適化処理のみで 10 倍もの性能向上が得られました。

さらに、OpenVINO™ ツールキットの DL Workbench を用いて FP32 から INT8 への量子化を行ったところ、実行時間は 0.59 秒となり、最適化からさらに 2.8 倍もの性能向上が得られています。今回は OpenVINO™ ツールキットのパフォーマンス優先の量子化を行いましたが、分類精度を示す指標 AUC の評価では、FP32 の平均 AUC=0.843 から INT8 で平均 AUC=0.842 となり、演算を軽量化

した場合でも精度の低下はわずかで、実用上の問題はないと考えられます。

加えて、各ステップが直列に実行されていた推論処理を見直し、8画像を非同期 (async) で実行するようにスクリプトを書き換えてチューニングを行ったところ、最適化適用においては 1.65 秒を 0.45 秒に、量子化適用においては 0.59 秒を 0.19 秒に、それぞれ短縮を図ることができました。

この 0.19 秒は他社の高性能な AI アクセラレーター・チップと比べてもほぼ同等の性能といえます。

医療分野では、医用画像の自動分類・注釈のほかに、電子カルテの言語処理などにディープラーニング技術を活用しようという動きが広まっています。普及のためには、医療現場の実情に即した機能や使いやすさはもちろん、導入と運用に要するコストをいかに抑えるかが重要です。

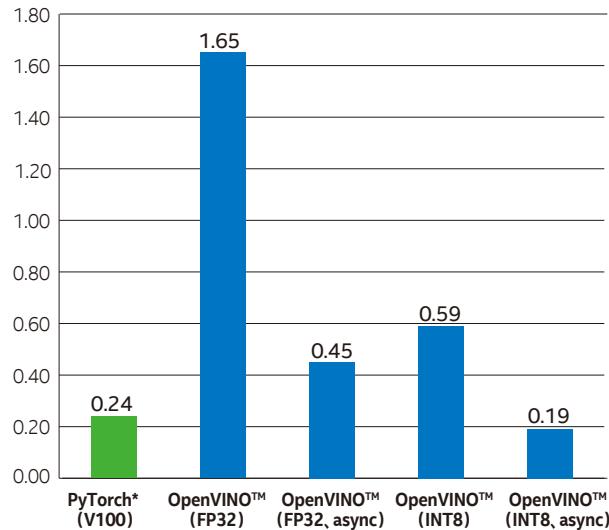
今回、OpenVINO™ ツールキットを用いた最適化と量子化を評価したところ、PyTorch* 経由で推論モデルをそのまま与えた場合に比べて、17.79 秒が 0.19 秒と、およそ 100 倍近い性能向上が得られました。

分類処理の性能要件を下げられれば、オンプレミス・サーバーの初期投資コストあるいはクラウドの運用コストを抑えられるほか、医師が日常的に使用している一般的なデスクトップ PC やノート PC でも処理が可能であり、実用化や普及の一助になると考えられます。

また、より高精度なディープラーニングの推論モデルによる読影技術が確立されれば、2020年初頭から世界で猛威を振るっている新型コロナウイルス感染症の胸部 CT 画像の診断にも役立つと期待されます。

図2. 推論モデル CheXNet* の実行性能の比較

Inference Performance (sec/iter, lower is better)



PyTorch* : 最適化および量子化なし、OpenVINO™ ツールキット FP32 : 最適化を適用、OpenVINO™ ツールキット INT8 : 最適化と量子化を適用、async : 画像分類処理の並列化。プロセッサーは 第2世代のインテル® Xeon® Gold 6258R プロセッサー、2.70GHz。

マルチモーダルでの疾患推定技術を確立し、医療サービスの向上に役立てていきたい。

国立研究開発法人理化学研究所 光量子工学研究センター データサイエンティスト 種石 慶 氏

ディープラーニングの推論処理における計算コストの課題に対して評価を行ったところ、OpenVINO™ ツールキットによる最適化と量子化によって大きな性能向上が得られることが分かりました。また、インテル® DevCloud については、Jupyter Notebook* およびバッチシステムが利用できるなど、技術者にとって使い慣れた環境が用意されていて、その上で OpenVINO™ ツールキットによる性能評価を効率的に実施できたと考えています。

インテル® DevCloud について

インテル® DevCloud は、ディープラーニングの推論モデルの開発、推論モデルのベンチマークリング、FPGA ロジックの設計、API のトレーニングなどを対象にしたリモートクラスター環境です。

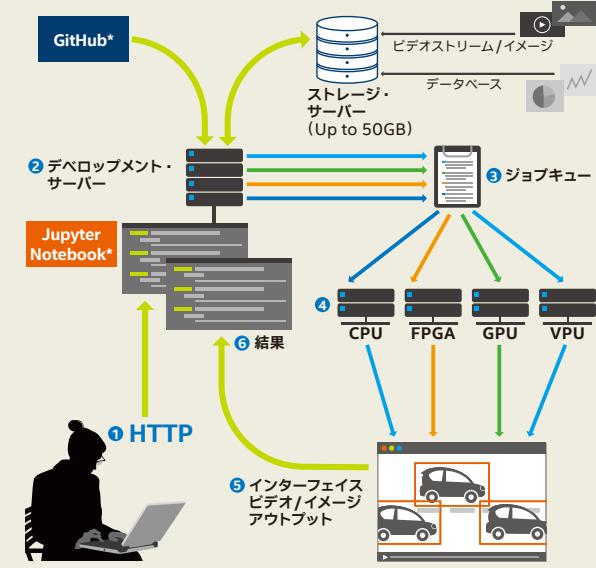
エッジ・コンピューティング向けの「DevCloud for Edge」、FPGA 開発向けの「DevCloud for FPGA」、および、データセンター系ワークロード向けの「DevCloud」の3種類のサービスがあり、それぞれの用途に応じて、インテル® Xeon® プロセッサー、インテル® Core™ プロセッサー・ファミリー、インテル® プロセッサー・グラフィックス、インテル® ニューラル・コンピュート・スティック 2、インテル® Movidius™ ビジョン・プロセシング・ユニット、インテル® Arria® 10 FPGA、インテル® Stratix® 10 FPGA などのハードウェア・リソースをリモートで使うことができます。

開発用ソフトウェアとしては、OpenVINO™ ツールキットのほか、ハードウェアを問わずに統合的なアプリケーション開発を実現するインテル® oneAPI、インテル® FPGA の開発ツールであるインテル® Quartus® Prime 開発ソフトウェア、および、オープンソースの PyTorch* や TensorFlow* などのディープラーニング・フレームワークが用意されています。

インテル® DevCloud の利用は最初の30日間は無料です。簡単な審査のち、お客様にアカウントが発行されます。詳しくはリンク先^{*4}を参照してください。

将来的には、医用画像のほか、血液検査データ、バイタルデータ、電子カルテデータなど、複合的な要素を組み合わせたマルチモーダルでの疾患推定技術を確立し、例えばリモート診療への応用を提案するなどして、医療サービスの向上にディープラーニング技術の応用を提案していきたいと考えています(図3)。

図3. インテル® DevCloud の動作の概要



[*1] CheXNet : <https://stanfordmlgroup.github.io/projects/chexnet/>

[*2] ChestX-ray14 : <https://nihcc.app.box.com/v/ChestXray-NIHCC>

[*3] OpenVINO™ ツールキット : <https://software.intel.com/content/www/us/en/develop/tools/openvino-toolkit.html>

[*4] インテル® DevCloud : <https://software.intel.com/content/www/us/en/develop/tools/devcloud.html>



この文書は情報提供のみを目的としています。この文書は現状のまま提供され、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、他者の権利の非侵害性、特定目的への適合性、また、あらゆる提案書、仕様書、見本から生じる保証を含みますが、これらに限定されるものではありません。インテルはこの仕様の情報の使用に関する財産権の侵害を含む、いかなる責任も負いません。また、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。

Intel、インテル、Intel ロゴ、Arria、Intel Core、Movidius、OpenVINO、Quartus、Stratix、Xeon 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。

インテル株式会社

〒100-0005 東京都千代田区丸の内3-1-1

<http://www.intel.co.jp/>

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

2020年10月