ホワイトペーパー

OpenVINO™ ツールキット

医用画像のマルチモーダル推論や化合物とタンパク質の相互作用予測を最新世代のプロセッサーが加速する

ディープラーニングを活用した医療データの解析や創薬技術の開発に取り組む理化学研究所の種石 慶 氏をお招きし、研究の概要や展望を伺いました。第 3 世代インテル® Xeon® スケーラブル・プロセッサー・ファミリー（開発コード名：Ice Lake）を使った性能ベンチマークについてもお話しいただきました。

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

医用画像のマルチモーダルな推論や創薬 AI に取り組む

大内山：本日は理化学研究所の種石 慶様をお招きしてディープラーニングを活用した研究の一端についてお話を伺いたいと思います。種石様にはインテルのフォーラムでたびたび講演していただいていますが、どのような研究に取り組んでいるのかあらためて教えてください。

種石：医療データ解析と創薬 AI の 2 つを大きなテーマとして研究を進めています。医療データ解析は 2014 年頃から手掛けていて、その代表的な 1 つがディープラーニングを用いた医用画像の認識や分類です。もう 1 つの創薬 AI については 2005 年頃から取り組んでいます。まだディープラーニングが知られていなかった頃で、機械学習の一手法であるサポート・ベクター・マシン（SVM）を使って、低分子化合物の活性予測などを行っていました。

大内山：医用画像への応用については「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構築 / 最適化や量子化で 100 倍の性能向上を確認」というホワイトペーパー[*1]で紹介しましたが、ほかにはどのようなモデルの研究を進めているのですか？

種石：現在は少し視点を変えて、眼底画像に血液検査を含む健診データを組み合わせた「マルチモーダル」な推論モデルの研究に取り組んでいます。眼底画像だけから血液検査でしか得られない情報を低侵襲で推定することができれば、予防医療などに役立てられるのではないかという考えです。

大内山：学習や推論の研究を進めるにあたってハードウェアの性能は研究効率にも大きく関係するかと思います。日ごろ使用している環境を教えてください。

種石：学習で主に使っているのが、2020 年 10 月から運用が始まった理化学研究所の「HOKUSAI SailingShip」（HSS）という大規模なデータ科学基盤です。HSS は第 2 世代インテル® Xeon® スケーラブル・プロセッサー・ファミリー（開発コード名：Cascade Lake）を搭載した 440 ノードのサーバーで構成されていて、総コア数は 21,120 コアです。この CPU ファームから必要に応じて計算ノードを創薬 AI の研究などに割り当てて使用しています。推論の実行では、性能ベンチマークも兼ねて、第 2 世代インテル® Xeon® スケーラブル・プロセッサー・ファミリーや第 3 世代インテル® Xeon® スケーラブル・プロセッサー・ファミリーを含めて複数のプラットフォームを利用しています。

胸部 X 線画像からの疾患推定
第 3 世代インテル® Xeon® スケーラブル・プロセッサー・ファミリーで 1.3 倍の性能向上を確認

大内山：前回のホワイトペーパー[*1]では、PyTorch で開発した推論モデルをそのまま実行した場合に比べて、インテルの OpenVINO™ ツールキットによるモデル最適化でおよそ 10 倍、INT8 への量子化を適用するとさらに 2.8 倍、といった性能向上が得られることを示していただきました。

使用されたハードウェア・プラットフォームは第 2 世代のインテル® Xeon® Gold 6258R プロセッサー（2.70GHz/28 コア）でしたが、その後、第 3 世代のインテル® Xeon® Gold プロセッサーでもベンチマークされたそうですね。

種石：胸部 X 線画像データベースである「CheXNet」（DenseNet-121）の画像を対象に疾患を推定するモデル[*2]をオンプレミス環境に置いた第 3 世代のインテル® Xeon® Gold 6330 プロセッサー（2GHz/最大 3.10GHz/28 コア）で実行してみたところ、クラウド環境であるインテル® DevCloud for the Edge 上の第 2 世代のインテル® Xeon® Gold 6258R プロセッサー（2.70GHz/最大 4GHz/28 コア）に比べて、FP32 では 1.54 秒が 1.19 秒に、INT8 では 0.38 秒が 0.35 秒に高速化されました（図 1）。

コア数は同じですが、コア周波数が違うこと、オンプレミスとクラウドの違いがあること、および、インテル® Xeon® Gold 6258R プロセッサーでの実行にはバージョン 2019R3 の OpenVINO™ ツールキットを用いたのに対して、インテル® Xeon® Gold 6330 プロセッサーでの実行にはその時点で最新の 2021.3 を用いましたので、プロセッサー性能の純粋な比較にはならないのですが、それでも FP32 の場合で、1.3 倍ほどの性能向上が得られたのはとても魅力的に感じます（図 1）。

大内山：当社の評価でも、第 3 世代インテル® Xeon® スケーラブル・プロセッサー・ファミリーは第 2 世代インテル® Xeon® スケーラブル・プロセッサー・ファミリーに比べて 1.4 倍から 1.7 倍程度の性能向上が得られることが分かっています。種石様が説明された 1.3 倍程度の性能向上は、それぞれのベンチマーク条件に違いがありますのでご指摘のように厳密な比較にはなりませんが、種石様が説明された 1.3 倍程度の性能向上は、プロセッサーの違いとして十分にリーズナブルな数値と考えます。

第 3 世代インテル® Xeon® スケーラブル・プロセッサー・ファミリーで推論モデルをなぜ高速に実行できるかというと、ベクトル演算を高速化する「AVX-512」命令を実行する際に発熱を抑えるためにクロック周波数を下げる仕組みが入っているのですが、第 3 世代インテル® Xeon® スケーラブル・プロセッサー・ファミリーは第 2 世代インテル® Xeon® スケーラブル・プロセッサー・ファミリーに比べて下げ幅を抑えているんです。合わせて、メモリー帯域幅の向上なども貢献していると考えています（コラム参照）。

低分子化合物とタンパク質の相互作用予測にグラフ表現や言語モデルを応用

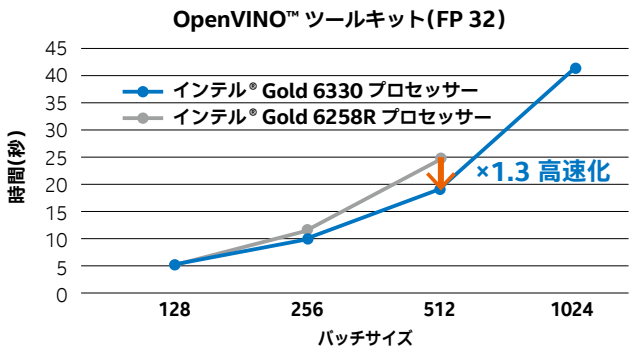
大内山：続いて創薬 AI のご研究について伺います。そもそも創薬にディープラーニングはどのように関わってくるのでしょうか。

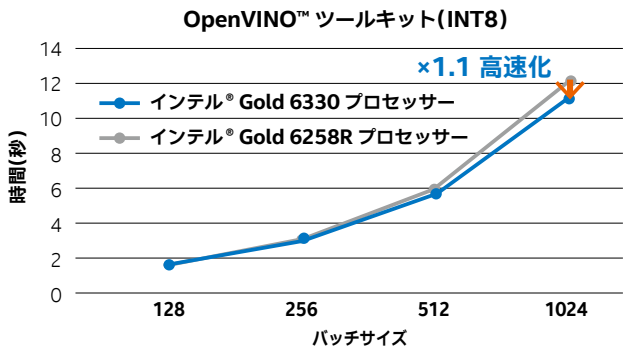
種石：創薬のアプローチも時代とともに変化しています。以前からがん細胞やウイルスなどに特異的な分子標的に作用する低分子化合物を探索する手法が注目されていますが、近年はがん細胞などの抗原に結合する抗体を産生する抗体医薬が 1 つのトレンドになっていますし、タンパク質の機能を分子動力学を用いてシミュレーションする取り組みも進んでいます。

そうした中で私が取り組んでいるのが、化合物とタンパク質との相互作用予測です（図 2）。分子標的薬の候補となる低分子化合物に対しては、元素をノード、二重結合などを特徴とした化学結合をエッジとするグラフ表現を用いたグラフ・ニューラル・ネットワーク（GNN）を使って、理論的には 10 の 60 乗以上の構造を持つ低分子化合物の活性、毒性、物性などを推定します。

タンパク質に関しては二次構造や機能の予測に取り組んでいます。タンパク質は 20 種類のアミノ酸の配列によって構成されていて、そのシーケンスは データベースに登録されているものだけでも数億種類にも及びます。タンパク質の機能を解析するには分子動力学シミュレーションを行うのが 1 つの方法ですが、計算コストが高いため、あらゆるタンパク質に適用するのは現実的ではありません。そこで一次的なスクリーニングとして、ディープラーニングを用いてタンパク

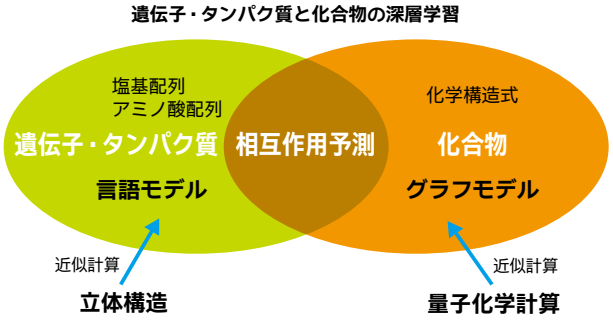
図 1. 胸部 X 線画像から疾患を推定するモデルのベンチマーク結果





・FP32 最適化モデル 平均 AUC=0.843 → INT8 量子化モデル 平均 AUC=0.842

図2. 創薬 AI 研究における化合物の活性予測と タンパク質の構造・機能予測の関係



質の構造や機能を推定しようという狙いです。長大なアミノ酸配列が対象となるため、推定には Transformer や BERT といった言語モデルを使います。

大内山：低分子化合物に適用している GNN は比較的新しいニューラル・ネットワークと思いますが、どのぐらいの推定精度が得られているのですか？

種石：低分子化合物の活性を推定する問題では、AUC（Area Under The Curve ：精度を面積で表した指標）で 0.95 から 0.97 ぐらいが得られています。一次スクリーニングに十分使えるレベルと考えています。

大内山：創薬 AI を対象にした推論のベンチマーク結果があれば紹介してください。

種石：OpenVINO™ ツールキットは使わずに PyTorch 1.7 で作成したモデルを、前述のように第 2 世代のインテル® Xeon® Gold 6258R プロセッサーと第 3 世代のインテル® Xeon® Gold 6330 プロセッサーで実行してみたところ、GNN の場合でバッチサイズが 1024 のとき、6258R が 134.26s だったのに対して、6330 ではおよそ 1.9 倍高速となる 69.57s が得られました（図 3）。[*3]

一方のタンパク質の二次構造を推定する BERT モデルでは、バッチサイズが 128 のとき、6258R では 1.21s だったのが 6330 では 1.7 倍高速の 0.71s となりました。[*4]

ちなみに GNN の場合、使われるメモリー容量はバッチ数の 2 乗に比例するため、大容量かつ高速なメモリーの搭載が高速化のポイントの 1 つになります。

ただし、TB オーダーの主メモリーを DDR4 で構成するのはコストがきわめて高くなってしまい現実的ではありません。そこで、インテルが主メモリーと外部ストレージのギャップを埋める新たなメモリー階層として提案している、容量あたりのコストが DDR4 よりも安い「インテル® Optane™ パーシステント・メモリー」[*5] を、インテル® Xeon® Gold 6330 プロセッサーのハードウェア・プラットフォーム 2.0（128GB×16）TB を搭載し、ベンチマークを行ってみました。

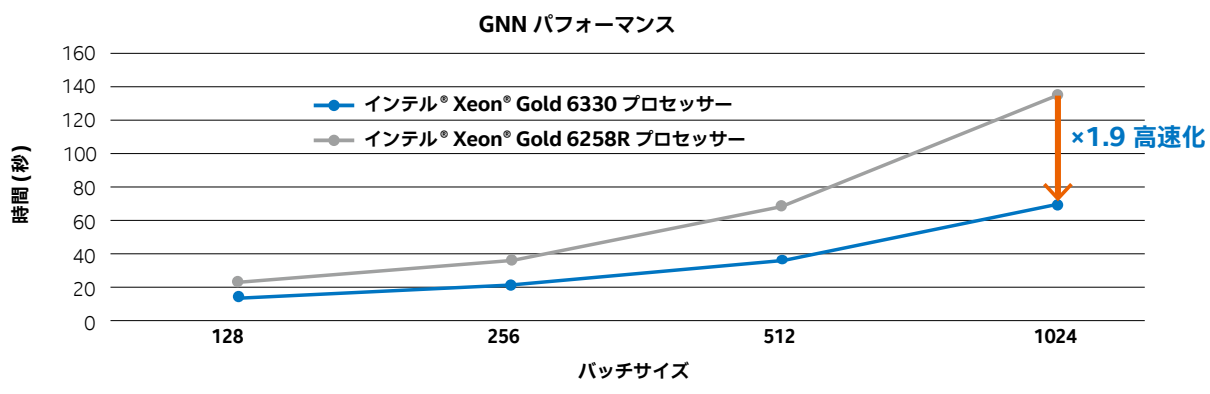
その結果、128GB の DDR4 のみで主メモリーを構成した場合には難しかった 4096 や 8192 といった大きなバッチサイズでも実行できることが確認できました。性能は DDR4 に比べれば若干低下してしまいますが、メモリーコストを抑えられることを考えれば十分に許容できる範囲と考えています。このように、インテル® Optane™ パーシステント・メモリーを活用することで、より大規模な GNN の計算ができるとの見通しを得ることができました。

大内山：第 2 世代から第 3 世代にするだけで、GNN で 1.9 倍、BERT モデルで 1.7 倍もの性能向上が得られたという種石様のベンチマーク結果はとても参考になりました。また GNN には、NAND 型フラッシュで構成されていて、DDR4 に比べて容量あたりのメモリーコストを最大で 40% も抑えられるインテル® Optane™ パーシステント・メモリーが効果的との考察も興味深く感じました。こうした結果をディープラーニングの研究や応用に取り組むお客様にご紹介していきたいと思います。

推論性能で鍵を握る実行環境のチューニング 第 3 世代インテル® Xeon® スケーラブル・プロセッサー・ファミリーや OpenVINO™ ツールキットの活用もポイント

大内山：ディープラーニングの研究や応用を進めていくうえで推論性能はとても重要と思いますが、プラットフォームの構築ではどのようなポイントに注意すべきでしょうか？

図3. 低分子化合物の活性予測を行う GNN モデルの実行性能比較



種石：データ量の増加、マルチモーダル化による情報量の増加、およびモデルの大規模化が進んでいますので、十分な性能を得るには推論の実行環境にもさまざまな工夫が必要です。

例えば、プロセッサーの各コアにワークロードを均等に分散させる、FP32 を bfloat16（BF16）に精度変換したり INT8 に量子化して計算量を抑える、データやパラメーターをできるだけ主メモリーに配置する、主メモリーと外部ストレージの間にインテル® Optane™ パーシステント・メモリーのような階層メモリーを設ける、マルチノードで並列処理を行う場合はインテル® Omni-Path アーキテクチャーのようなスケーラブルなインターコネクターを用いる、といったポイントが挙げられるかと思います。

また、今回紹介したように、モデルの最適化や量子化にはインテルの OpenVINO™ ツールキットが有効ですし、基本性能が向上している第 3 世代インテル® Xeon® スケーラブル・プロセッサー・ファミリーの活用も検討するに値すると考えます。

大内山：ありがとうございます。種石様のような先進的なユーザーの継続的なフィードバックによってインテルのソリューション開発は支えられています。今後もご研究にご活用いただければ幸いです。最後に研究の展望などをお聞かせください。

種石：バイオ・テクノロジーの分野ではゲノム編集技術に代表されるイノベーションが進んでいます。それに伴い、患者個人に合った医薬品を生成的に設計しようというニーズも生まれています。ディープラーニングの学習モデルはそのための有力な手法の 1つですが、一方でタンパク質はアミノ酸の配列が 1つ変わるだけでも機能が変化することがあるので、大規模化が進む言語モデルを使ってそういった複雑性を表現できるようになるのか、実のところまだよく分かっていません。

ただ、タンパク質や遺伝子を対象にしたディープラーニング技術の進展は、新たな医薬品の開発のみならず、生命現象の解明にも大きく貢献すると期待されますので、引き続き研究を進めていきたいと考えています。

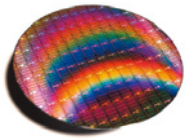
大内山：これからのご研究に期待しています。本日はありがとうございました。

第 3 世代インテル® Xeon® スケーラブル・プロセッサー・ファミリー

インテルは 2021年4月に、データセンター向け最新プロセッサーとして、第 3 世代インテル® Xeon® スケーラブル・プロセッサー・ファミリーを発表しました[*6]。内部アーキテクチャーのコードネームから「Ice Lake+」とも呼ばれています。製造のプロセスノードは 10nm です。

第 2 世代インテル® Xeon® スケーラブル・プロセッサー・ファミリーに比べて、クロックあたりの処理命令数を 1.2 倍、メモリー帯域幅を最大 1.6 倍、メモリー容量最大を 2.66 倍、PCI レーン数を最大 1.33 倍にそれぞれ増強しており、パフォーマンスは平均で 1.46 倍向上しています。

また、ディープラーニングの推論モデルの実行を高速化する「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）」アクセラレーション機能の強化も図られていて、第 2 世代インテル® Xeon® スケーラブル・プロセッサー・ファミリーに比べて 1.74 倍程度の推論性能の向上を実現しています



*1 https://www.intel.co.jp/content/dam/www/public/ijkk/jp/ja/documents/solution-briefs/riken-using-dl-to-infer-disease-from-x-ray-images-openvino-paper.pdf
*2 https://github.com/taneishi/CheXNet/
*3 https://github.com/taneishi/GNN_molecules/
*4 https://github.com/taneishi/ProtTrans/
*5 https://www.intel.co.jp/content/www/jp/ja/architecture-and-technology/optane-dc-persistent-memory.html
*6 https://www.intel.co.jp/content/www/jp/ja/products/docs/processors/xeon/3rd-gen-xeon-scalable-processors-brief.html
本ホワイトペーパー記載のベンチマークは理化学研究所の種石 慶氏が独自に実施されたものであり、性能比などはインテルが公表している値と一致しない場合があります。

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/

2021年8月

348067-001JA
JPN/2108/PDF/CB/BCG/YS