学習済みAIモデルを活用する

◆AIの利用はエッジデバイスなどさまざまなデバイスに急展開している

ものづくりの現場で製品検査にAIが導入され、品質向上など大きな影響を与えている。当初クラウドで運用していたAIシステムもデータ量の増大がネックになり、エッジデバイスへのAI搭載が加速している。従来のAIモデルでは、適用するシステムやサービスごとに求められる計算の精度や消費電力などの性能が異なり、モデルサイズや演算量、学習用データを使った検証などを人がそれぞれ試行錯誤しながら時間をかけて設計・開発していた。

学習済みAIモデルを演算量に応じ使い分けるスケーラブルAIの技術開発や、評価された学習済みAIモデルを公開するなど、AI開発を支援する動きが出ている。

◆学習済みAIモデルを有効利用し開発効率を上げる技術の開発

2021年8月20日、東芝と理化学研究所は、学習済みAIモデルの性能をできるだけ落とさず、演算量が異なるさまざまなシステムに展開することを可能にするスケーラブルAIを発表した。システムのAI展開における課題解決が期待される。

従来、スマートフォンや監視カメラ、無人搬送車（AGV）に人物検出AIシステムを構築する場合には、それぞれAIモデルを開発・学習する必要があった。

スケーラブルAIは、元となる学習済みのフルサイズの深層ニューラルネットワーク（フルサイズDNN）と、それぞれのアプリケーションに応じ演算量を削減したコンパクトDNNからなる。これまでのコンパクトDNNは、演算量削除時に、すべての層で単純に行列の一部を削除したが、重要な情報が多い層の行列をできるだけ残すことで、近似による誤差を低減させた。また、学習時にそれぞれのDNNの出力値と正解との差が小さくなるようにフルサイズDNNの重みを更新して調整することで、あらゆる演算量のモデルでバランスよく学習できる（図.1）。

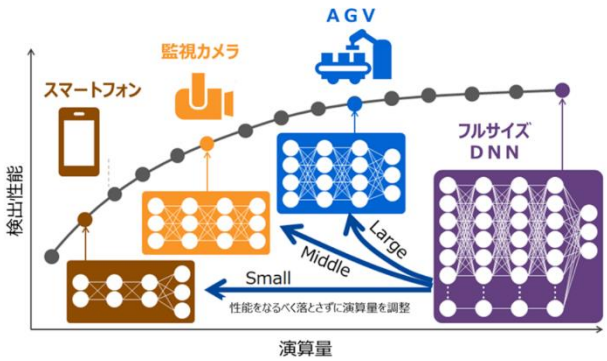

図１．スケーラブルAIの概念図

出典：東芝

世界的に知られている一般画像の公開データ「ImageNet」を用いた、被写体に応じてデータを分類するタスクの精度評価によると、フルサイズDNNから演算量を2分の１、3分の１、4分の１に削減した場合、分類性能の低下率は、従来のスケーラブルAIが2.7%、3.9%、5.0%だったのに対して、新技術ではそれぞれ1.1%、2.1%、3.3%に抑えることができた。今後ハードウエアに対する最適化を進め、さまざまな組み込み機器やエッジデバイスへの適用を進める。

◆MLCommonsは機械学習のトレーニング性能の評価結果を公開した

21年6月30日、MLCommonsは機械学習のトレーニング性能評価MLPref Training v1.0の結果を公開した。MLCommonsはGoogle、NVIDIA、IntelなどAIのハードウエア、ソフトウエア、システムに関連する50社によるコンソーシアムで、機械学習のイノベーションを加速させるため、データベースの管理と開発環境の提供、AIモデルの品質測定の活動をしている。毎年、機械学習の性能をトレーニングと推論に分けて、それぞれに計測可能な複数のベンチマーク結果を18年から公開している。今回公開したトレーニングでは、画像分類、物体検出、音声認識、言語モデル、レコメンデーションなど8種類の評価基準について、精度、速度、効率の観点からベンダーの現在の性能を比較できる。また、データベース生成の方法やソースも公開されているため、AI開発者にとって効果的な支援活動となっている。

表　MLPerfの機械学習ベンチマーク項目

領域	評価基準	データセット	モデル名	目標品質	実行数	トレーニング内容	アプリケーション
視覚	画像分類	ImageNet	ResNet-50 v1.5	75.90%分類	5	入力画像に、決められた一連のカテゴリからラベルを割り当てる	自律走行車など
視覚	医療用画像セグメンテーション	KiTS19	3D U-Net	0.908平均DICEスコア	40	画像から腎臓の癌細胞を見つけてセグメント化する	MRI画像など
視覚	軽量物体検出	COCO	SSD	23.0%mAP	5	画像や動画内の顔、自転車、建物など実在するインスタンスの物体の周囲にバウンディングボックスを指定	物体の位置推定と種類分け
視覚	重量物体検出	COCO	Mask R-CNN	0.377ボックス最小APおよび0.339マスク最小AP	5	対象画像の個別のオブジェクトを検出し、ピクセルマスクを識別する	一般物体検出、姿勢検知など
言語	音声認識	LibriSpeech	RNN-T	0.058単語誤り率	10	大きなバッチ単位でRNN-Tの損失関数（目標と実際の出力誤差）の重み調節	スマホ上の会話の書き起こしアプリなど
言語	自然言語処理	ウイキペディア2020/01/01	BERT	0.72マスク-LM精度	10	ひとかたまりのテキストの中のさまざまな単語間の関係を使用してテキストを認識	質問への回答、文の言い換えなどの言語関連
商業	レコメンデーション	1TBクリックログ	DLRM	0.8025 AUC	5	専用トレーニングソフトでユーザーと製品や広告などのサービスアイテムとのやり取りを理解させる	検索機能の結果のランク付けやクリック率の
リサーチ	強化学習	囲碁	Mini Go	チェックポイントに対して50%の勝率	10	19×19の盤面で対局する囲碁を使用して、さまざまな手を評価し、戦略的効果を最大に高める	囲碁

出典：https://mlcommons.org/en/training-normal-10/　よりARC作成

今回は13社がベンチマークを実施しており、DELLやNVIDIAのように8項目すべての項目に結果を公開している企業もあれば、特定の項目に絞っている企業もある。過去の結果と比較することで、AIモデルの改善状況の確認や、AI導入時の、処理速度に見合うシステム性能の目安として活用できる。

学習済みAIモデル活用で、AIの社会実装が加速することを期待する。【成田誠】