組み込みディープラーニングがもたらす新しい多様な可能性

マーティン・カッセル

無数のプロセッサやチップセットによって、組み込みビジョンシステムへのディープラーニングの実装が可能になり、ネットワークの負荷の軽減や遅延の低下がもたらされる。

オンデバイスでディープラーニング推論を実行できる組み込みビジョンシステムは、組み立て、協働ロボット、医療技術、ドローン、運転支援、自動運転など、多種多様な業界の用途に対する有効なソリューションとなる。ディープラーニングを組み込みデバイスに実装することの、クラウドベースのソリューションと比べた場合のメリットとしては、ネットワークの負荷の軽減と遅延の低下があり、それは新しい用途の開拓につながる。ただし、それを行うには、正しい組み合わせのツールとハードウエアが必要である。

開発者は、利用可能なハードウエア、処理リソース、メモリのサイズと種類、速度要件（イメージ周波数、処理、帯域幅など）を考慮に入れる必要がある。その他の検討項目として、必要な画像分解能または画質、消費電力、システムサイズなどがある。このようなシステムに対し、スマートカメラ、ビジョンセンサ、シングルボードコンピュータ（SBC）といった一連の組み込み画像処理システムが提供されている。

ディープラーニングにおいて、ニューラルネットワークのトレーニング（学習）は組み込みデバイス上では実行されない。デバイス上で直接行われるのは、学習済みネットワークの実行（推論）である。完全に学習済みのネットワークは、一般的に1つの特定の作業（表面検査など）のみを対象とするが、さまざまな種類の組み込みシステムで実行することができる。これを行うには、学習済みネットワークをOpen Neural Network Exchange（ONNX）やNeural Network Exchange Format（NNEF、http:/bit.ly/VSD-NNEF、**図1**）などの特殊フォーマットや、共有重み付きのネットワーク記述ファイルに変換する。多数のニューラルネットワークの中でも、畳み込みニューラルネットワーク（Convolutional Neural Network：CNN）は、他のネットワークよりも重みが少ないことから高性能かつ低消費電力であるため、ディープラーニングに特に適している。そうした特長は、CNNがほとんどの組み込みビジョン用途に対して実行可能である理由でもある。

リソースを節約するために、圧縮（ディープラーニングアルゴリズムやデータの圧縮）やプルーニング（特定の特性、ニューロン、重みなど、結果に小さな影響しか与えないネットワーク部分の削除）などの手法を適用して、CNNを軽量化することができる。また、演算精度を8ビット、あるいは4ビットの固定小数点にまで落とすことによって（量子化）、CNNを簡素化することも可能である。バイナリニューラルネットワーク（Binarized Neural Network：BNN）は、2値の重みを使用し、ニューラルネットワーク層における固定小数点乗算を1ビット演算にすることで、さらなる軽量化を行うものである。BNNは、必要な処理能力、消費電力、サイクルレートが低いが、その代わりに演算精度が低すぎるので、組み込み用途には適用できない可能性がある。

精度を落とすことの影響は、用途に

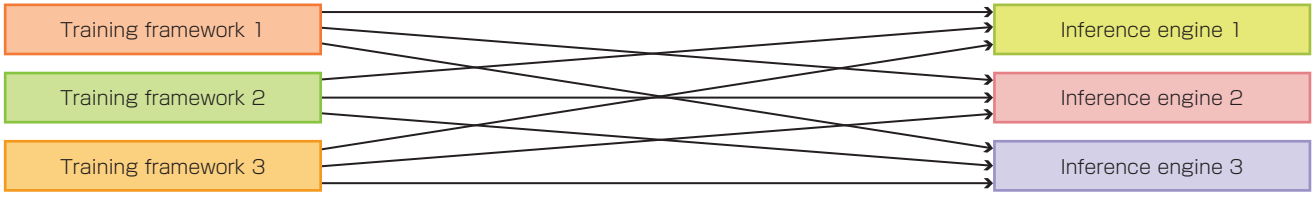

図1 Neural Network Exchange Format（NNEF）により、ハードウエアメーカーは、学習フレームワークと推論エンジンの間で、学習済みニューラルネットワークを確実に受け渡すことができる。

よって異なるが、必要なストレージ容量を少なくする効果はある。Trained Ternary Quantization（TTQ）は、重みを1度に2ビットに格納される3値に変換する手法である。重みは各層ごとに、ゼロ（無駄な接続を切断するために用いる）、正値、負値の3つの値に量子化される。組み込みシステムでは最終的に、あらゆる場合における性能と重みの軽量化の間の適正なバランスを図る必要がある。

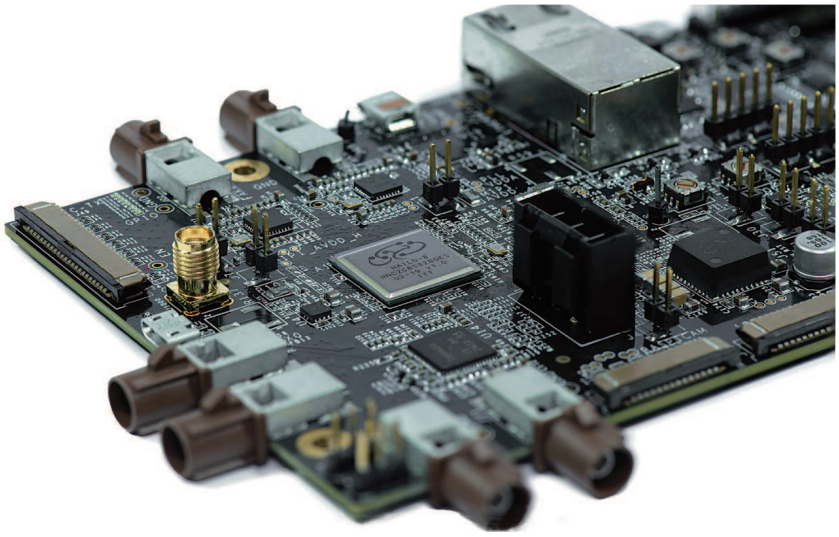

図2　事前学習済みのニューラルネットワーク「ResNet-50」を、8ビット、224×224の分解能で実行する「Hailo-8」プロセッサは、1.7Wの消費電力で672fpsのフレームレートと、2.8TOPS/Wの演算性能を達成する。

適切な手法の選択

ディープラーニングにおいて、CNNは、物体検出またはパターン／異常識別の手法を適用して画像を処理し、例えば分類などの結果を出力する。組み込みシステムは、大量のデータを処理するという点において、PCに劣る。そのため、CNNを実行する組み込みシステムには、5～50W程度の最小限の消費電力で、高い演算能力（5～50 TOPS[Tera Operations Per Second]）とそれに見合う大きな帯域幅が必要である。システム性能は、消費電力要件と、顧客がシステムに対してどれだけの金額を支出するつもりがあるかによって決まる。

用途や、データ量の要件に応じて、中央処理ユニット（CPU）と組み合わせてアクセラレータとして機能するさまざまなチップセットが、提供されている。例えば、米インテル社傘下の米モビディウス社（Intel Movidius）によるビジョン処理ユニット（VPU）である「Myriad 2」及び「Myriad X」や、イスラエルのハイロ社（Hailo）によるディープラーニング専用プロセッサ「Hailo-8」（**図2**）などがある。このようなチップセットは、開発者やデバイスメーカーがディープニューラルネットワークや人工知能（AI）を組み込みデバイスに実装できるようにすることを目的に、特別に設計されている。

組み込みシステムにおいて、CPUは通常、ARMアーキテクチュアをベースとする。CPU単体では並列構造を持たず、十分な演算能力が得られないため、推論を実行するにはプロセッサの追加が必要になる。一方、グラフィックス処理ユニット（GPU）は大規模な並列性と広いメモリ帯域幅を備え、メインメモリの負荷を軽減する。GPUは熱出力も高く、演算速度を高めた組み込み版として最適化されている。例としては、米エヌビディア社（NVIDIA）の「Jetson」や、米AMD社の「Ryzen Embedded」などがある。

ある顔認識の実験では、CPUによる認証件数が毎秒1～4件だったのに対し、新しい組み込みGPUを使用した場合は最大400件にものぼった。この性能差を緩和して、少しでもGPUに近づけるために、CPUメーカーは既に演算速度の向上に取り組んでいる。

高い演算速度に低い熱出力と低い遅延を併せ持つFPGA（Field-Programmable Gate Array）は、ディープラーニングの代替プロセッサとして大きな可能性を秘めている。例えば、開発者はFPGAをソフトウエアのようにプログラミングすることによって、さまざまなニューラルネットワークを実行するように変更することができる。経時とともに複数のニューラルネットワークを必要とするアプリケーションには、FPGAが適切な選択肢となる。

もう1つの選択肢であるASIC（特定用途向け集積回路）は、例えば、高速行列乗算や直接畳み込み用のエンジンによって、ディープラーニングアクセラレーション向けに最初から設計されている。ASICは、演算能力が高く、熱出力が低い。メモリアクセスを最小限にして、チップ上のデータ量を最大限に保つことにより、処理の高速化とスループットの向上が図られている。しかし、ASICはごくわずかしかプログラムできないため、実装の柔軟性の面でFPGAに劣る。また、個々のASICの製造コストは高い。それでも、組み込みビジョンシステムに適した選択肢の1つであり、ASIC同士を接続することも可能である。

以上まとめると、ディープラーニング用の組み込みビジョンシステムは、小さな処理ボードと、独バスラー社（Basler）の「dart BCON for MIPI」開発キット（**図3**）のような極小のカメラモジュールで、一般的に構成される。このアーキテクチュアにおいて、SoC（System on a Chip）はメイン処理ユ

図3　バスラー社の「dart BCON for MIPI」開発キットは、カメラモジュール、「Snapdragon」処理ボード、レンズやケーブルなどのアクセサリで構成される。

ニットとして機能し、アプリケーションアクセラレータとしてCPUを搭載し、GPU、FPGA、ディープラーニングチップセットといったその他のプロセッサを追加することができる。SoM（System on Module）には、メモリ（RAM）や電源管理といった重要な要素を追加して、組み込みディープラーニングに対して実用的な状態にしたSoCが含まれている。このようなシステムには通常、顧客が特定の要件に合わせて開発できるカメラなど、周辺デバイス用の物理的なコネクタを備えたキャリアボードも必要である。一方、SBC内のSoCは最初からキャリアボードに搭載され、周辺デバイス用のポートは固定である。また、組み込みOSは個々のコンポーネントを制御する必要がある。

ロボットの要件

固定式の産業用ロボット、または可動式の協働ロボットは、3Dイメージング技術を利用して周辺環境を認識する場合が多い。そうしたロボットは、物体を分類し、異常を検出し、他の人間や機械と事故を起こすことなく協働し、ツールの位置決めを行い、デバイスを組み立てる。ロボットが実行する多くの処理に対し、ディープラーニングは新たな次元の柔軟性を加える。

例えば、転移学習（transfer learning）では、事前学習済みのニューラルネットワークを複数のロボット業務に対して、時間的にもコスト的にも効率良く直ちに使用することができる。そうしたネットワークは、ロボットの動作を促進することによって深層強化学習（deep reinforcement learning）を使用し、新しい環境に適応する。報酬と罰（マイナスの報酬）については、ロボットが良い行動を繰り返すようにパラメータを変更することにより、デバイスの組み立てのような、可変要素を含む難しい検査環境においても、ロボットが比較的短時間で新しい作業段階を学習できるようにする。

変更されたロボット動作は、従来型のアルゴリズムでプログラム可能だが、かなりの労力が求められる。一方、ディープラーニングを使用すれば、ビンピッキングロボットは、位置や向きに関係なく部品を識別して取り上げることができる。部品が傾いていたり、他の部品に覆われていたりしても、問題ない。ロボットは自律的に自らの向きを変えて、ノーマル、ステレオ、3D画像といった入力データに基づいて既知の部品を検出する。

適切なターゲット用途としては、包装やパレタイジング、機械のロード／アンロード、ピックアンドプレース、ビンピッキング、自動車製造の品質検査、電子部品の組み立て、精密農業、作業ステージの自動化、医療技術（スマートデバイス、疾病の早期検出、手術支援システムなど）が挙げられる。

結論

ディープラーニングを組み込みデバイスに実装すると、クラウド内の保護画像の処理に伴うセキュリティリスクが緩和される。組み込みアプリケーションでは、データの生成と使用が一カ所で行われるため、プラント全体の効率が高まる。演算リソースの増加を、より強力な小規模ネットワーク、組み込みアプリケーション用の新しいプロセッサ、圧縮、プルーニング、量子化などの改良プロセスを使用することによって、埋め合わせることができる。

ディープラーニングに適したチップセットや処理ボードは、アプリケーションによって大きく異なり、個々のアプリケーションに応じてソフトウエアとともに正確に調整する必要がある。転送学習や強化学習などの技術を適用することにより、使用するネットワークと、それに伴うアプリケーション全体をすばやく変更して、製造の柔軟性を高めることができる。組み込みデバイス上のディープラーニングは、ロボット以外の分野でも主要な役割を担い続けることになるだろう。どのコンポーネントやネットワークがどのアプリケーションに適しているかという問題は、今後もシステムインテグレータにとっての課題となる。

著者紹介
マーティン・カッセル（Martin Cassel）は、独Basler社Silicon Software部門のテクニカルライター。
URL：http://www.baslerweb.com

VSDJ