ディープラーニングモデル最適化によるArmマイクロコントローラにおけるAIの開放

2020年4月21日 Mary Bennion

*****このブログ内のすべてのコンテンツは、Deeplite.aiの共同設立者でCPOであるDavis Sawyerが提供しています*****

組み込み機器やプラットフォームにおけるAIとディープラーニングの出現によって、さらにインテリジェントな製品開発を目的としたエキサイティングで新しい方法が生み出されています。コンピュータビジョンや自然言語などの分野では、ディープニューラルネットワーク（DNN）は複雑なタスクを実行するためのデファクトツールとなっており、画像の中の物体を認識する能力では人間よりも優れています。そのため、近年、 DNNは非常に複雑になり、計算量が増大し、セマンティック セグメンテーションや顔認識のような、より興味深くインテリジェントなユースケースを実行できるようになりました。そのため、多くの最先端のモデルアーキテクチャは、日常的な機器では実用的ではありません。現在、何十億ものマイクロコントローラ（MCU）が一般的に使用されており、結局のところAIが機器で使用されることが妨げられています。

ディープラーニングにサイズの問題があることは周知の事実です。例えば、言語処理用の大規模なトランスフォーマーモデルであるMegatronLMは、80億以上のパラメータ（33GBのメモリー）を持ち、トレーニングに500台のV100 GPUを9日以上必要とします。極端な例ではありますが、最新のDNNモデルでは処理可能なリソース要求も、電話、自動車、センサーなどのエッジデバイスで広く使用されている低消費電力のコンピューティングハードウェアでは処理できません。

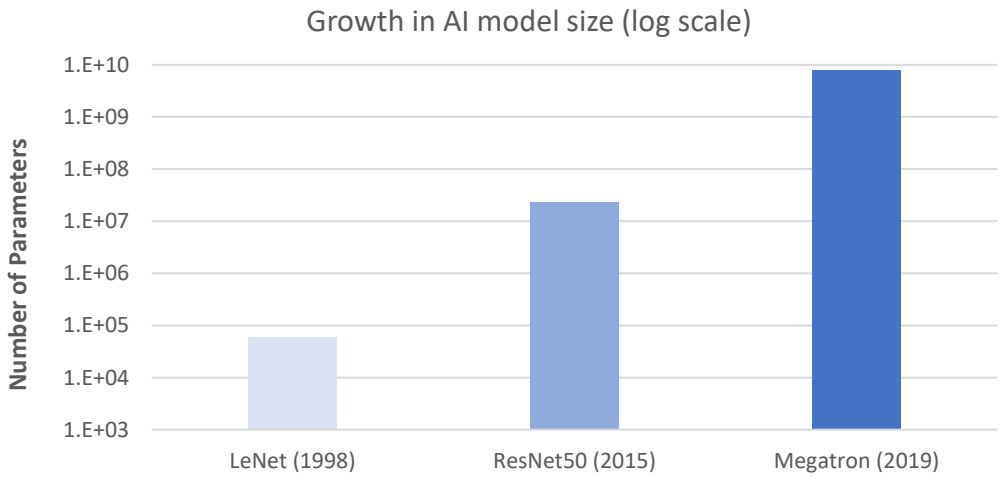

図1：ディープラーニングモデルの複雑性の拡大

ディープラーニングのチームにとって、最初の焦点はユースケースで高い精度を得ることができるモデルを作ることです。精度の高い結果を出すようにモデルを訓練することが望まれます。次に、モデルのサイズ、レイテンシ、消費電力について考慮します。これらの制約は使用可能なハードウェア次第であり、元のモデルで達成できた精度レベルを落とすことなく結果を出

すのは非常に困難です。デバイスに適用可能なターゲットハードウェア上でディープラーニングを実際に使用するには、複数の設計メトリクスを考慮しなければならないため、ディープラーニングモデルの生産展開に隔たりが生じています。

幸いなことに、解決策はあります。これらの障壁は、効果的なディープラーニングモデルの最適化によって解決することができます。モデルの最適化により、AIエンジニアは非常にコンパクトで高性能なモデルを迅速に作成できるようになります。それだけでなく、AIチームは、MCUのArmCortex-Mのような非常に効率的なプロセッサをバッテリー駆動でリソースが限られた機器向けに追加することもできます。モデル最適化により、従来はクラウド接続とサーバークラスのハードウェアに依存していた重要なリアルタイムタスクを、標準のMCUでローカルに実行することが可能になり、スループットと推論レイテンシの大幅な改善を実現できるようになりました。

次に、モデル最適化の影響を実証するため、スマートファクトリーの設定における次の展開を考察します。品質管理用に製品サンプルのポジ画像とネガ画像を分類するため、低消費電力カメラ、Arm MCUおよび畳み込みニューラルネットワーク（CNN）を使用します。

- Arm Cortex-M4（チップメモリー 256KB、フラッシュ 1MB）
- 独自のバイナリ クラシフィケーション データ トレーニングを実施したMobileNetv1 CNN
- ターゲット最適化基準は、モデルサイズ（MB）と精度です。

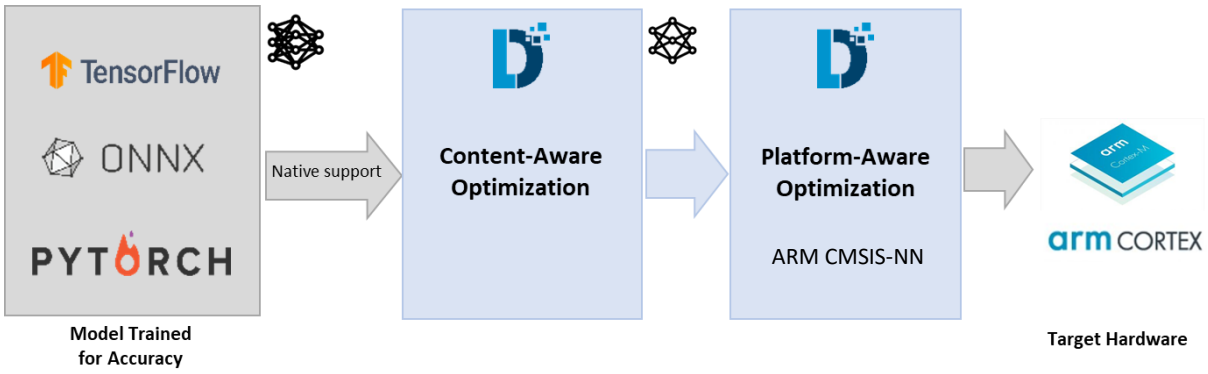

図2：ARMハードウェアのDNNデプロイメントにおける最適化ワークフロー

今日、ディープラーニングモデルを最適化する最も一般的な方法は、試行錯誤というコストのかかるプロセスを経ることです。エンジニアは、プルーニング、ハイパーパラメータ調整、量子化を行う必要があります。大抵、それらテクニックを組み合わせて、ハードウェア用の商業的な実用型モデルを見つけるために数週間から数カ月を費やします。Deepliteでは、自動の簡単な最適化プロセスで、DNNモデルの従来の型を変化させています。ユーザーは、事前に訓練されたモデル、データセット、いくつかの制約（例えばサイズや精度）を参照し、実行を押すだけです。当社のオンプレミスのソフトウェアエンジンは、独自の設計空間調査アルゴリズムを使用して効率的に集中し、デプロイメントの特定の制約に合わせて最適化された新しいモデルアーキテクチャを見つけます。以前は手作業で数週間から数カ月かかっていた作業を、シンプルで使いやすい1つのソフトウェアエンジンで数時間から数日でできるよう自動化しました。こ

れは、製造プロセスにとって非常に重要です。なぜなら製品の品質を検査し保証するため、個別に多段階のアプローチを必要とするからです。多くの場合、欠陥はプロセスの最後まで認識されませんが、DNNモデルの高速スループットによる自動光学検査は、早期検出（以下のスマート製造プロセスのステップ2で描かれているように）を可能にします。

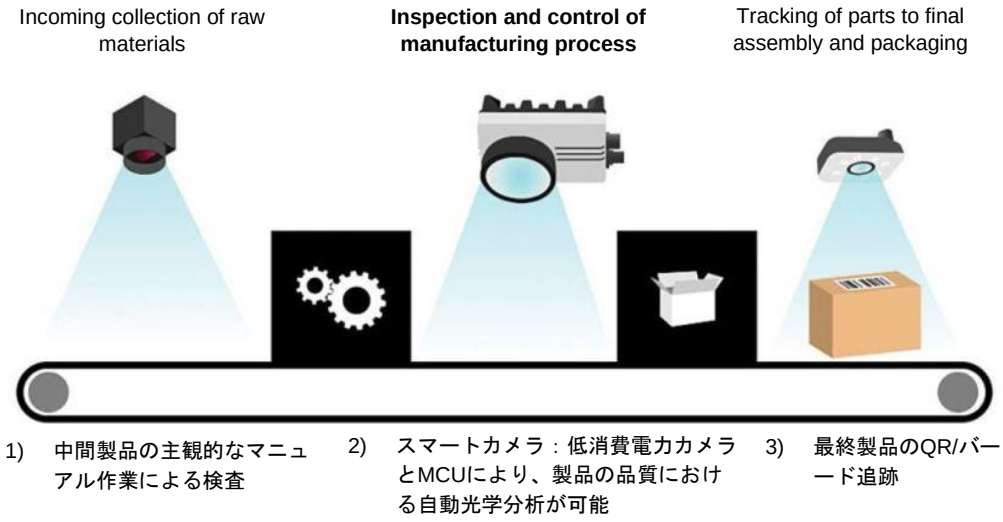

図3：スマート製造プロセス用のパイプライン自動検査。

Deepliteを利用することで、モデル最適化のための多目的アプローチにより、エンジニアは精度に焦点を当て、推論用に生産準備モデルをシームレスに作成することができます。上記のステップ2に注目すると、初期のMobileNetv1 モデル（約12.8 MB、精度92%のバリデーションデータ）は、ArmCortex-M 4を搭載した低消費電力カメラで実行する必要があります。しかし、今日までAIエンジニアはTop-1精度が2%未満、256 KBのオンチップメモリにパラメータが適合するモデルを作成する必要がありました。このような精度保持とサイズ縮小を達成するために、プルーニング、ハイパーパラメータ調整、および基本的な量子化をカットすることはできません。なぜなら、ほとんどの場合、これらのテクニックでは、実際のタスクを実行可能なソリューションを見つけるために必要なネットワーク設計スペース（フィルタの数、レイヤー、操作、カーネルサイズ）を十分に調査していないためです。業界のアプリケーションには、繰り返し使用可能で再現可能なアーキテクチャ検索の方法が必要なのです。Deepliteのエンジンは、独自のアテンションメカニズムで、事前にトレーニングされたモデルアーキテクチャを解析します。また、適切なソリューションを見つけるために必要な設計スペースを大幅に削減する有意義なセンシティビティとネットワークトランスフォーメーションを特定します。さらに、知識の蒸留を適用することにより、高忠実度および重要なユースケースの精度を大幅に維持することができます。 私たちのエンジンは、新しいネットワーク設計上の決定論的アプローチを用いて迅速に収束することができます。ユーザーが定義した制約に基づいて設計スペースを削減し、新しいコンパクトなネットワークを構築します。上記のケースでは、Deepliteのエンジンは、約144 KBの新しいアーキテクチャを自動的に検出し、Top-1の精度を1.84%削減するだけで、MAC操作の数を大幅に減少することができます。

さらに、初期の大規模モデルを訓練し、 Deepliteの最適化エンジンを使用することで、最適化後のモデルに正則化の効果を与えることができるため、目に見えないデータにおける一般化能力が高いことが分かりました。デプロイメントに必要な基準を超えることができ、ユーザーがArm-Cortex M4でローカルにクラシフィケーションのモデルを実行できるようになりました。また、MCUで直接推論を実行することにより、レイテンシ、帯域幅、推論コストのようなキーエリアでの著しい削減を達成することができます。

さらにDeepliteは、PytorchやTensorFlow、ONNXといったAIフレームワークや、Arm NNやCMSIS-NNといったローレベルツールと相互運用可能です。

これにより、設計チームは簡単にモデルを新しいハードウェアに移植したり、既存のコンフィギュレーション内にとどめることができます。Armとのパートナーシップにより、Deepliteのディープラーニングモデル最適化へのアプローチは、Arm Cortex-M 4のような非常に効率的なハードウェアを、AIチームがAIタスクにエッジデバイスで活用することを可能にしています。

生産準備モデルアーキテクチャの開発サイクルを自動化することで、Deepliteは市場投入までの時間を大幅に短縮し、多目的最適化によってモデル開発チームの生産性を向上させることもできます。最後に、モデルの最適化をCortex-M 3、M 4、M 55などのArmの低消費電力ハードウェアと組み合わせることで、ユーザーはスループットとエネルギー節約において、多くの利益を得ることができます。

モデル	サイズ（バイト）	GMac	パラメータ（100万）	精度
初期	12836104B	0.583	3.21	Top1:92.443%
最適化後	**144186B (89.04x)**	**0.112 (5.21X)**	**0.14 (22.26X)**	**Top1:90.607%(-1.84%)**

表1：独自のデータでのMobileNetv 1の最適化結果の要約。

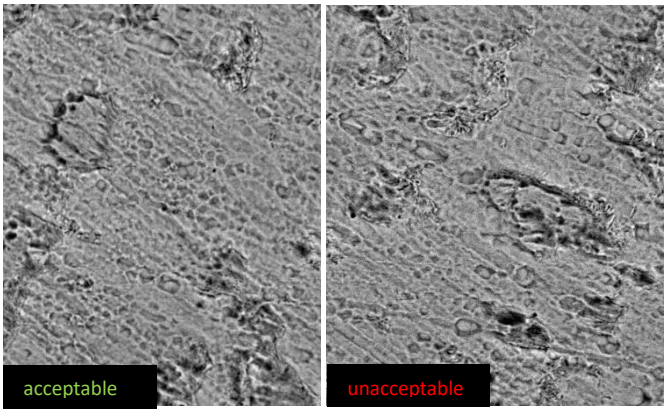

図4：品質検査用のサンプルバイナリクラシフィケーションイメージ

高度に最適化されたモデルと新しいAIのユースケースの結果は、次世代のインテリジェント製品に急速に展開されています。Deepliteは、スマート製造、自動車、消費者向け機器業界の指導

者に、従来の限界を超えてAIモデルを展開し、AIチームを今まで到達したことのない領域に導いています。DeepliteとArmは一体となって、MCUでのデバイス推論を可能にすることで、エッジIoTの可能性を解き放つべく取り組んでいます。詳細またはお問い合わせはこちらへ：https://www.deeplite.ai/.