エッジAIの変革
最新マイコンに搭載された
ニューラル・ネットワーク
処理ユニットの性能

はじめに

人工知能（AI）のユビキタス化が進み、生活や社会の環境の中で収集するデータの量が指数関数的に増加しています。こうした中、すべてのデータをクラウドで遠隔処理することは持続不可能であり、非現実的になってきています。さまざまな製品やアプリケーションでAIを広く活用していくためには、より効率的でローカルな処理ソリューション、すなわちエッジAIの導入が不可欠になります。ソリューションの開発者は、ニューラル・ネットワーク処理アクセラレータを組み込んだ先進的マイクロコントローラを活用することで、エッジ・デバイス上でAIのパワーを直接活用することができるようになります。このアプローチは、消費電力の削減、ネットワーク負荷の低減、レイテンシの低減など、大きなメリットをもたらし、より高速で応答性の高いAI駆動型アプリケーションを実現します。

どこでもAIの時代が到来

人工知能は、従来のルールベースのアルゴリズムやハードコードされた命令によるコンピューティングの限界を打ち破り、テクノロジーとの関わり方に革命をもたらしています。AIによって、機械が自律的に問題の解法をデータから学習し、新しい入力への適応や不確実性な事象を処理することができるようになりました。

パターンの認識や例外の検出、予測、新しい未知のデータを扱うための普遍化といったAIの技術は、さまざまなアプリケーションで採用されています。例えば、スマートフォンの写真撮影、フィットネス用のウェアラブル・デバイスによる装着者の活動検出、スマート・リテールなどの目的での人数カウント、スマートシティでの交通カメラや群衆管理、セキュリティ・システム、機器の予知保全などで積極活用されています。

AIは、一般的に使用されているさまざまなテクノロジーを、より賢く、より効率的に、より迅速に変えることを可能にします。

- **パーソナル・エレクトロニクス:** スマートフォン、スマートウォッチ、ホーム・アシスタントは、よりインテリジェントになり、パーソナライズされた体験や予測機能を提供するようになりました。
- **車載AI:** 自律走行車、先進運転支援システム（ADAS）、スマート交通管理の開発に不可欠な情報処理手段になりました。
- **産業応用でのAI:** 予知保全を強化し、サプライチェーンを最適化し、製造や物流における業務効率を向上させています。
- **ヘルスケアAI:** 診断、個別化医療、患者モニタリングに革命をもたらしています。
- **小売業用AI:** 販売状況などの分析とレコメンデーション・システムが、ショッピング体験を一変させています。

データ取得場所のより近くで活用

大規模なデータセットを使用したAIアルゴリズムのトレーニングには、ニューラル・ネットワークを対象にした大規模な並列処理と多層データフロー処理を実行可能なAIサーバの演算能力の活用が不可欠です。これらの処理は、従来、クラウドで行われてきました。また、訓練されたアルゴリズム、つまりAIモデルもクラウド・サーバ上にホストして活用していました。

あらゆるハイテク機器において、AIシステムの搭載が急速に求められるようになりました。こうした状態が顕在化している現在、クラウドではなく、それぞれの機器に直接AIモデルをホスティングすることで、いくつかの利点が得られるようになってきています。車両ナンバー・プレートの読み取り、ビデオ・フレーム内に映っている人を自動的に検出・カウントすることで、システムがイベントに反応するまでの時間を大幅に短縮できるようになりました。クラウド接続を通じて共有するデータ量も削減できるため、ネットワーク帯域幅の需要を軽減できる効果も期待できます。また、自律型デバイスでは、ネットワークが停止した場合でも動作を継続可能です。さらに、ネットワーク上で共有するデータ量を最小限に抑えることで、データのプライバシーを保護することも可能になります。そしておそらく持続可能性の観点から最も重要になる利点は、AIのワークロードを効率的な低消費電力デバイスに移行することで、AIデータ・センターで現在消費しているエネルギーを劇的に削減できることです。より一般的呼称としてエッジAIとして知られるこのオンデバイス推論は、組込みシステム、特にアプリケーションをホストする処理サブシステムの設計アプローチを変革しつつあります。

ABIリサーチは、エッジAIの市場が、今後10年間で大きく成長する見込みであると予測しています。このデータは、農業、自動車、セルラー・ネットワーク、ヘルスケア、製造、個人用および作業用デバイス、小売、ロボットなど、さまざまな分野で**エッジAIアプリケーション用マイクロコントローラ・ユニットの利用が大幅に増加**することを示唆しています。同市場は指数関数的な成長を遂げ、**2030年には18億ユニット**近くに達すると予想されています。製品開発者がAIで設計を強化するために取り組まなければならない主な課題は以下の通りです。

- **エネルギー需要/消費:** AIや機械学習（ML）のタスクは計算量が多く、消費電力が大きくなります。これは、エネルギー効率が極めて重要な、電力制約のある環境ではAIの活用が特に困難になることを意味しています。
- **性能のボトルネック:** 汎用マイコンは、ニューラル・ネットワーク（NN）の処理に要する演算性能への要求を満たすのが困難になりがちです。そして、AIアプリケーションの効果的活用を妨げる性能ボトルネックが発生します。
- **レイテンシとリアルタイム処理:** 多くのAIアプリケーション、特にリアルタイムの応答を必要とするアプリケーションにとって、長い待ち時間は重大な懸念事項になります。レイテンシが長いと、AI駆動システムの性能やユーザ体験が損なわれる可能性があります。
- **開発での複雑さとコスト:** マイクロコントローラ・ユニットにAIアルゴリズムを組み込むためには、これらのデバイスのメモリ容量や帯域幅、処理能力が限られているため、いくつかの課題を抱えることになります。さらに、利用可能なソフトウェア・ツールやAIフレームワークは、もともと組込み開発者を念頭に置いて設計されていないことが多く、そのことが原因で開発プロセスを複雑にしています。

ニューラル・ネットワーク処理ユニット（NPU）とは何か?

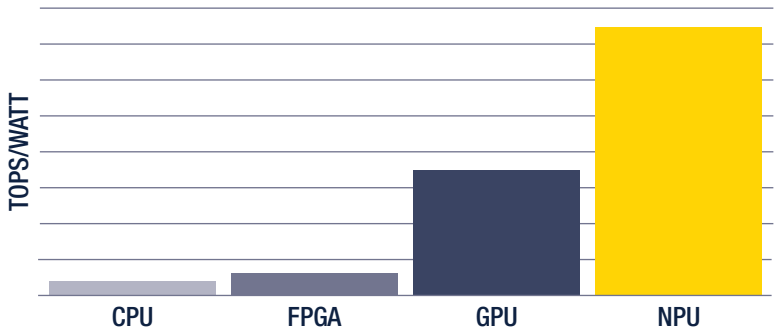
一般的な組み込みプロセッサ・コアは、命令・フェッチ・デコード・実行といった演算に伴う一連の内部処理を逐次実行に最適化して設計されています。AIモデルを、可能な限り効率的に実行するようには設計されていません。ニューラルネット・ワークの計算トポロジでは、従来アーキテクチャでは最適化されていない、莫大な量の積和演算や乗算や頻繁なメモリ・アクセスを行います。一般的な組み込み機器の消費電力とチップ面積に関する制約の中で、高速かつ効率的にAI推論を実行できる別のアーキテクチャが求められています。ニューラル・ネットワーク処理ユニット（NPU）は、こうした要求に応えるために登場しました。

NPUを中央演算処理装置（CPU）やグラフィックス・プロセッシング・ユニット（GPU）と比較することで、AIやMLアプリケーションにおけるNPUの明確な優位性を理解することができます。以下の表は、それぞれの主な特徴と違いを示しています。

特徴	CPU	GPU	NPU
主要機能	汎用処理	グラフィックスのレンダリング処理、並列処理	ニューラル・ネットワークの加速
アーキテクチャ	少数の強力なコア、高クロック	多数の小型コア、SIMDアーキテクチャ	ニューラル・ネットワーク専用コア
処理方向	逐次処理	並列処理	並列処理と低遅延処理
命令セット	複雑な命令セット（x86、Armなど）	グラフィックスと並列計算命令	畳み込みニューラル用に最適化
エネルギー効率	中程度	中～高	高い
ユースケース	一般的なコンピューティング、制御タスク	グラフィックスのレンダリング処理、科学シミュレーション、MLトレーニング	エッジAI、リアルタイム推論、IoTデバイス

ニューラル・ネットワークに最適化した特殊アーキテクチャを導入することによって、NPUは、エッジAIアクセラレーションにおいて大きな進歩を遂げています。そして、従来のプロセッシング・ユニットよりも優れた効率を実現可能になりました。

図1：各種ハードウェア・アーキテクチャの電力効率



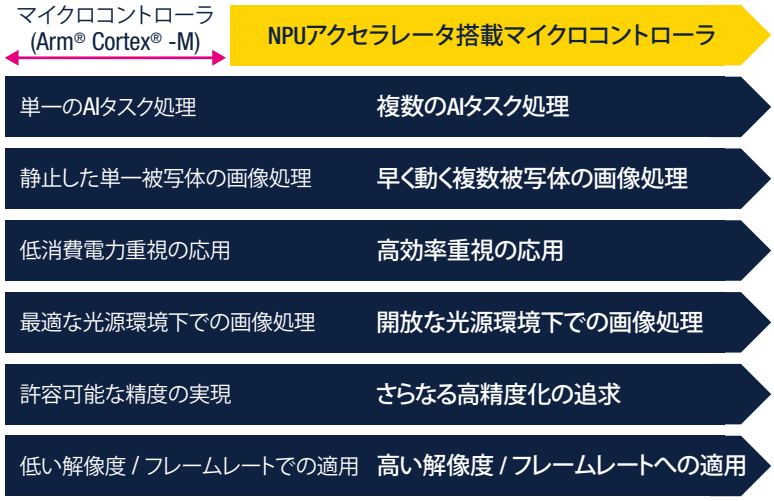
GPU: グラフィック・アクセラレータ
NPU: ニューラル・ネットワーク処理ユニット（AIアクセラレータ）

Source W. Dally

組み込みAIアプリケーションの新たな可能性を拓く

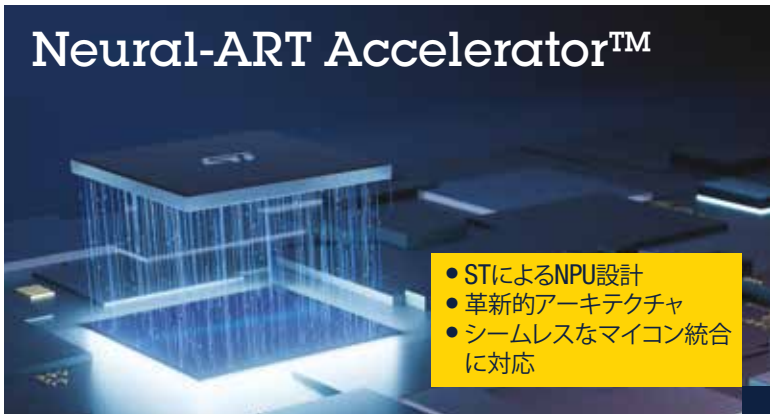
NPUは高効率であるため、マイクロコントローラ・ベースの組み込みアプリケーションのような厳しいエネルギー制約のある環境への適用に、特に適しています。低消費電力を維持しながら、幅広い組み込みAIのユースケースに対応する最適なソリューションを提供します。

図2: 組み込みAIの新たな可能性を拓く



NPUとマイコンを統合することによって、マイコンの能力が大幅に拡張され、これまで手が届かなかったより複雑なAIタスクを処理できるようになります。従来、マイコンは処理能力やエネルギー効率に制約があるため、適用可能なAIアプリケーションが、低解像度の画像解析や時系列解析、低フレームレートといった単純なものに限られていました。しかし、NPUを追加することで、マイコンによって、より高速に移動し、より小さな対象物に対して、音声認識、物体分類、姿勢推定、それぞれの物体の位置特定などの高度なAI機能を実行できるようになりました。AI推論タスクをNPUにオフロードすることで、マイコンは他の重要な機能に集中することが可能になり、効率的かつリアルタイムでの処理が可能になります。

ST Neural-ARTアクセラレータ™ ST独自のニューラル・ネットワーク処理ユニット

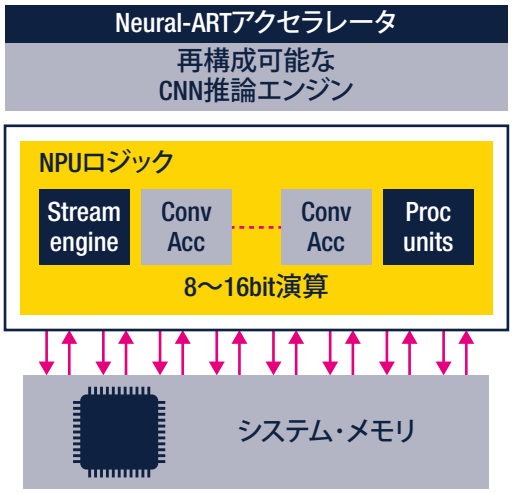


STでは、AIをデバイスに直接組み込み、エッジAIを実現する未来を構想しています（組み込みAI）。エッジでAI機能を実現することによって、クラウド・コンピューティングへの依存度を低減し、エネルギー消費の削減とレイテンシの低減を目指します。こうしたビジョンを実現するため、STは**Neural-ARTアクセラレータ™**シリーズの最初のバージョンである、高度に並列化したハードウェア・コプロセッサを搭載した専用設計のNPUを発表しました。STM32マイクロコントローラに統合されたこの画期的なNeural-ARTアクセラレータは、組み込みデバイス上でのAI推論タスクの効率的処理を可能にします。

本製品は、よりスマートでエネルギー効率の高いソリューションを提供するというSTの責務に沿って、組み込みAIの実用化と普及を推進することを目指した製品です。この新世代のNPUには、STのマイコン技術とAIにおける広範な専門知識が注がれており、多様なアプリケーションに向けて、強力かつ効率的で拡張可能なソリューションを提供します。

NPU自体と同様に重要なのが、開発者がアクセラレータのアーキテクチャを効率的に利用する際に、モデルの評価、最適化、および実装を可能にする付属の開発ツールチェーンです。ST Edge AI Suiteでは、開発者がKeras、TensorFlow、ONNXといった一般的AIフレームワークを活用して、STM32マイコン向けのAIアプリケーションを構築するためのツールを提供します。

図3：Neural-ARTアクセラレータ™のアーキテクチャ概要



Neural-ARTアクセラレータのアーキテクチャ概要

Neural-ARTアクセラレータ™は、さまざまな推論カーネルに対応可能な複数の専用ハードウェア・アクセラレータを統合しています。これらのアクセラレータは、再構成可能なデータフロー・ストリーム処理エンジンを介して動的に接続されており、柔軟で効率的な処理を実現します。このアーキテクチャには、以下のような構成可能な固定小数点MACを備えた、設定可能な数の畳み込みアクセラレータが導入されています。その精度は16bitまたは8bitです。

ソフトウェア・ツールとの統合

Neural-ARTアクセラレータ™ IPの高い効率性は、STM32Cubeソフトウェア・エコシステムとのシームレスな統合によって確保されています。この統合は、**STM32Cube.AI**（デスクトップ・ツール）および**ST Edge AI Developer Cloud**（オンライン・ツール）によってネイティブにサポートされています。これによって、NPUの機能を完全に活用するための最適化されたコードを手間なく生成可能にしています。

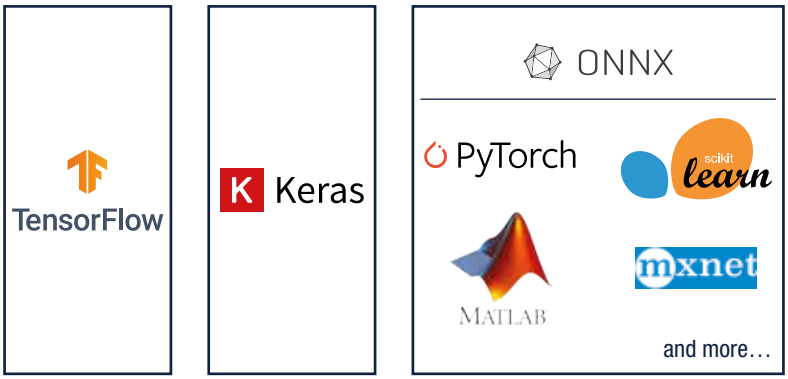
STM32Cube.AIまたはST Edge AI Developer Cloudを利用することによって、開発者は、事前にトレーニングされたニューラル・ネットワーク・モデルを、Neural-ARTアクセラレータ上で効率的に動作するコードへと簡単に最適化して変換することができます。どちらのツールも、ニューラルネット・ワークを分析し、その演算子を適切なハードウェア・リソースにマッピングし、各層が最適な方法で高速化されるようにします。こうした包括的ツールチェーンを活用することによって、STM32マイクロコントローラへのAIモデルの導入が簡素化され、開発者はハードウェア・アクセラレーションの複雑さに煩わされることなく、イノベーション創出に集中できます。

まとめると、Neural-ARTアクセラレータは、マイコン・ベースのアプリケーションに高度なAI機能を組み込むための強力でエネルギー効率の高いソリューションを提供します。**STM32Cubeソフトウェア・エコシステム**との統合によって、スムーズで効率的な開発プロセスを実現し、開発期間を短縮します。

図4：組み込みAIソフトウェアのエコシステム

ニューラル・ネットワーク（NN）モデルを開発する際、AI演算子はモデルが実行するさまざまな数学的・論理的関数を定義する上で重要な役割を果たしています。TensorFlow、Keras、ONNXなどの各AIフレームワークでは、開発者がモデルの構築と学習に活用可能な演算子のセットを提供しています。これらの演算子をソフトウェアでサポートすることはAIアプリケーションを実現するために不可欠であり、ハードウェアのサポートは最大の性能を実現するために不可欠です。各フレームワークで利用可能な演算子の数は、開発プロセスに大きな影響を与えます。演算子のカバー範囲が広ければ、エッジAIアプリケーションの開発に必要な時間が大幅に短縮されるからです。

先進的ハードウェアと堅牢なソフトウェア・ツールを組み合わせることで、**ONNXを含む一般的なAIフレームワークから130以上の演算子**をサポートし、包括的な機能と柔軟性を提供します。



AI処理性能のベンチマーク

以下の表に示すデータは、一般的な組み込み機械学習（ML）モデルを処理する場合に、Neural ARTアクセラレータ™とArm® Cortex®-M55（逐次処理のみ）を活用したSTM32を、Cortex-M55単独と比較した推論アクセラレーションのベンチマーク結果を示しています。ただし、Cortex-M55には、機械学習およびデジタル信号処理（DSP）アプリケーションのパフォーマンスを大幅に向上させるMプロファイル・ベクター拡張を提供するArm Heliumテクノロジーが導入されており、AI推論にも最適化されていることをあらかじめ指摘しておきます。

表1：Neural-ARTアクセラレータ™ Gen1により、4畳込み配列処理を1GHz動作設定で測定した結果

	Cortex-M55のみを使った推論処理@400MHz		ST Neural-ARTアクセラレータによる推論処理@1GHz		アクセラレータによる改善効果
使用モデル	時間（ms）	fps	時間（ms）	fps	
MobileNet v1[1]	2244	0.45	19.4	51.54	x116
MobileNet v2[2]	1385	0.72	21.1	47.45	x66
Tiny Yolo v2[3]	3895	0.26	30.6	32.71	x127
Yolo v8n 256[4]	1821	0.55	31	32.26	x59
Yamnet 1024[5]	252	3.97	9.8	101.7	X26

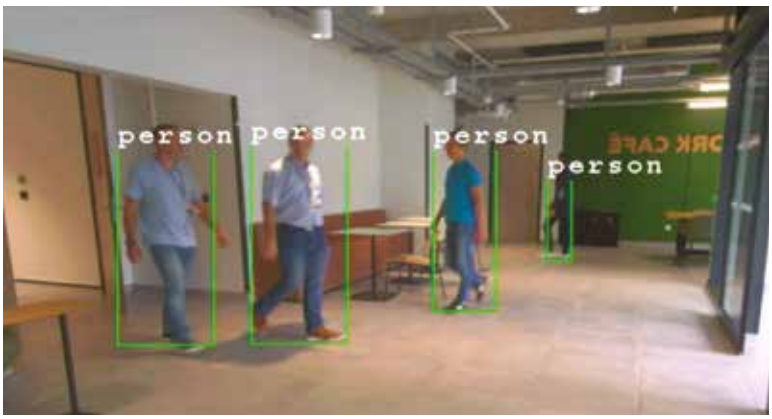
注記
1 **画像分類**：量子化したint8、入力解像度224x224x3、ImageNetデータセットで学習、モデルのフットプリント：4.45 MBの重みパラメータ、1.53 MBの実行サイズ
2 **画像分類**：量子化したint8、入力解像度224x224x3、ImageNetデータセットで学習、モデルのフットプリント：4.14 MBの重みパラメータ、2.08 MBの実行サイズ
3 **物体の検出**：量子化したint8、入力解像度224x224x3、COCOデータセットで学習、モデルのフットプリント：10.55 MBの重みパラメータ、0.38 MBの実行サイズ
4 **物体検出**：量子化したint8、入力解像度256x256x3、COCOデータセットで学習、モデルフットプリント：3.05 MBの重みパラメータ、1.6 MBの実行サイズ
5 **音声イベントの分類**：量子化したint8、入力解像度64x96、AudioSetデータセットで学習、モデルのフットプリント：3.4MBの重みパラメータ、0.14MBの実行サイズ

コンピュータ・ビジョンおよびオーディオ信号アプリケーションにおけるSTのNeural-ARTアクセラレータの性能に関するベンチマーク結果を見ると、組み込み型NPUが現在開発者に提供する利点が明確にわかります。MobileNet、Yolo、およびYamnetは、コンピュータ・ビジョン、物体検出、およびオーディオ信号イベントの分類に使用される一般的なAIモデル・トポロジです。**STM32Cube.AI**は、32bitマイクロコントローラに搭載されているArm Cortex-M CPUコア上でこれらのモデルをコンパイルするために使用しました。同じモデルをNeural-ARTアクセラレータ用にコンパイルし、実行時間とサポートフレームレートの対応する数値を記録しました。表には性能を比較し、得られた改善効果を示しています。

加えて、2つのユースケースを評価し、Neural-ARTアクセラレータによって実現した性能向上が、スマート・ビルディング、スマート・リテール、スマートシティなどの領域で広く使用されているマシン・ビジョン・アプリケーションにおいて、リアルタイムの物体検出と特徴認識をいかに効果的に実現できているのか検証しました。

最初の例は、人物検出と追跡アプリケーションへの適用を想定した物体検出モデル、Yolo（You Only Look Once）v8の例です。比較表では、Neural-ARTアクセラレータによって、Yoloの性能を100倍以上向上できていることを示しています。この例では、システムは26fpsのフレーム・レートを実現しています。この結果は、Yolo v8の推論によって、ビデオ・フレームの100%をリアルタイムで連続的に分析できることを示しています。同様の分析は、Cortex-M55でYolo v8を実行した場合には不可能です。

ケース・スタディ1：小売店などでの顧客追跡アプリケーションを想定した人物検出

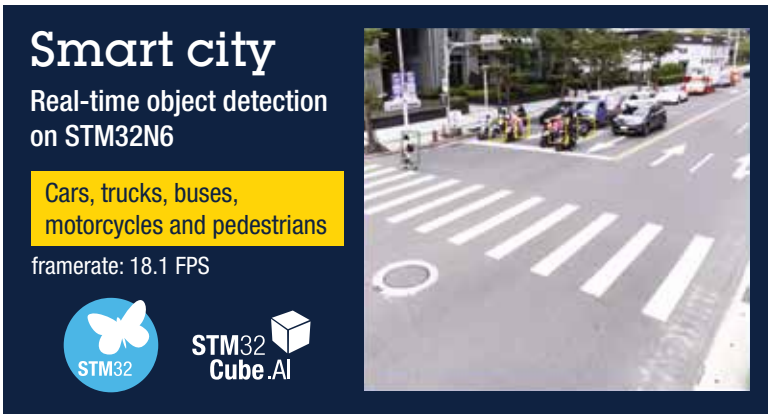


人物検出・追跡
Yolo v8推論320x320
Neural-ART with 4 Convolution Array（@1GHz = **26 fps**）

実際のデモ・ボード
画面キャプチャ

次の例は、スマートシティ・カメラが、小規模なTinyYolo v2を使用して、前の例よりも高い解像度で街頭シーンを解析するものです。この場合では、道路を走行する自動車の車種分類と歩行者検出に使用する画像解析への適用を想定しています。アプリケーションでは、18fpsのフレーム・レートにより、道路の安全性と交通の流れを監視し、改善するのに十分な情報を収集することができます。

ケース・スタディ2: スマートシティでの道路状況の監視アプリケーション



マルチクラスオブ物体検出、追跡、カウント
TinyYoloV2 416x416
Neural-ART with 4 Convolution Array (@1GHz = **18fps**)

実際のデモ・ボード
画面キャプチャ

これらの例において、入力解像度は重要な要因となります。カメラ・センサと光学システムを変更することで解像度を上げることにより、モデルに送る情報が増えれば、任意のモデルの実行時間が長くなります。このため、実現可能なフレーム・レートは事実上低下します。逆に解像度を下げれば、より高いフレーム・レートを実現できます。ケース・スタディ2のスマートシティの例では、カメラの解像度を上げれば、カメラからの距離が遠い位置にある車両のナンバー・プレートを認識することができるようになります。この場合でも、性能を過度に低下させることはありません。一方、より広い視野角を使用すれば、監視対象エリアをカバーするカメラの設置数を減らし、システム全体の設置コストを削減することができます。

Neural-ARTアクセラレータ™を搭載したSTM32マイコン上での組み込みAIソリューションの開発と活用

STは、ST Neural-ARTアクセラレータ™を搭載したSTM32マイクロコントローラ上でのAIモデルの開発と展開をサポートする包括的なツールおよびリソースを提供しています。これらのツールは、開発プロセスを合理化し、開発者が最適な性能と効率を達成できるように設計されています。

STM32Cube.AIは、開発ワークフローにおける基盤となる重要ツールです。開発者は、事前にトレーニングされたニューラル・ネットワークモデルを最適化し、STM32マイコンおよびNeural-ARTアクセラレータ用のコードに変換することができます。また、幅広いAIフレームワークをサポートしており、開発者がすぐに使い始められるように詳細なドキュメントとチュートリアルを提供しています。

ST Edge AI Model Zooは、STM32マイコン向けに最適化された事前学習済みAIモデルのコレクションです。画像分類、物体検出、音声認識など、さまざまなアプリケーションに対応したモデルを用意しています。ST Edge AI Model Zooは、STM32マイコンにすぐに導入できるモデルを開発者に提供することで、開発時間と労力を大幅に削減します。

STでは、開発者が迅速に開発を開始できるように、ST Neural-ARTアクセラレータを使用したAIソリューションの実装方法を示すリファレンス・デザインとサンプル・コードを提供しています。これらのリソースには、詳細な回路図、コード例、アプリケーション・ノートなどが含まれており、AI開発のための貴重なアイデアとベスト・プラクティスを提供します。さらに、**ST Edge AI Developer Cloud**によって、開発者は任意のSTM32上でニューラル・ネットワークの性能とフットプリントを簡単にベンチマークできます。

サードパーティのサポートとエコシステム

ST Neural-ARTアクセラレータ™は、AIソリューションの開発と進歩に貢献する、**サードパーティ・パートナーによって開発されたさまざまなエコシステム**の恩恵を活用して利用することが可能です。このエコシステムには、AI開発のための追加ツール、ライブラリおよびサポートを提供するソフトウェア・ベンダ、研究機関、および業界の専門家が含まれます。

このエコシステムを活用することで、特殊なハードウェア・コンポーネント、高度なAIアルゴリズム、専門家のガイダンスなど、幅広いリソースにアクセスすることができます。サードパーティのパートナーと協力することによって、開発プロセスを加速し、最高水準の性能と信頼性を満たすAIソリューションの開発が可能になります。

開発ワークフロー

ST Neural-ARTアクセラレータの開発ワークフローによって、開発者は、STM32マイクロコントローラ上にAIソリューションを迅速に実装できます。このプロセスには、いくつかの重要なステップが含まれます。

- **モデルの選択とトレーニング:** 開発者は、特定のアプリケーションに適したAIモデルを選択または設計することから始めます。このモデルは、通常、TensorFlow、Keras、またはONNX形式のモデルなどの高レベル・フレームワークを利用して、適切なデータセットを使用してトレーニングします。
- **モデルの最適化:** モデルの学習が完了したら、STの統合されたパートナー・エコシステムで利用可能なソリューションを使用して、量子化、プルーニング、圧縮などのさまざまな手法によって最適化することができます。
- **STM32Cube.AIによる変換:** STM32Cube.AIを活用することで、開発者は、事前にトレーニングされたニューラル・ネットワーク・モデルを最適化されたCコードへと変換することができます。このツールは、一般的なAIフレームワークからのモデルのインポートをサポートし、最適化技術を適用し、性能と効率を最大化するためにNPUコンパイラと統合します。
- **統合とテスト:** 生成したコードは、STM32マイクロコントローラ上で動作するアプリケーション・コードに統合されます。開発者は、**STM32CubeIDE**またはその他の互換性のある開発環境を使用して、ターゲット・ハードウェア上でアプリケーションをコンパイル、アップロード、テストできます。
- **デプロイメント:** アプリケーションを徹底的にテストして検証したら、エッジ・デバイスに実装できます。ST Neural-ARTアクセラレータは、AIモデルの効率的な実行を保証し、リアルタイムな性能と低消費電力を実現します。

結論

さまざまな分野でAIが急速に採用されていることから、AIアルゴリズムをマイクロコントローラ・ユニットに組み込む際の消費電力、性能の制限、レイテンシ、複雑性といった課題への対処が極めて重要になってきています。冒頭で強調したように、クラウドでAIの推論を処理するために必要なエネルギー消費は持続不可能であり、効率的でリアルタイムのデータ処理を保証するために組み込みAIへの移行が求められてきています。

Neural-ARTアクセラレータは、このような課題に対処する上で大きな進歩をもたらします。ニューラル・ネットワーク処理ユニット(NPU)をSTM32マイコンに統合することによって、Neural-ARTアクセラレータは、エッジ・デバイス上でのAI推論タスクの効率的処理を可能にします。この統合により、クラウド・コンピューティングへの依存度が下がり、AI処理を実行する際に消費するエネルギーが節約され、待ち時間が短縮されます。さらに、異常検知、音声認識、物体分類などの複雑なAI機能を実行するマイコンの機能が拡張されます。

専用ハードウェア・アクセラレータと再構成可能なデータフロー・ストリーム処理エンジンを備えたNeural-ARTアクセラレータのアーキテクチャは、高い性能と柔軟性を実現します。**STM32Cubeソフトウェア・エコシステム**とのシームレスな統合によって、開発プロセスはさらに強化され、開発者は事前にトレーニングされたニューラル・ネットワーク・モデルを、NPU上で効率的に動作する最適化されたコードへと簡単に変換することができます。高度なハードウェアと堅牢なソフトウェア・ツールの組合せにより、Neural-ARTアクセラレータは飛躍的な性能向上を実現し、幅広いアプリケーションに理想的な選択肢となります。

性能のベンチマーク結果から、Neural-ARTアクセラレータは、従来のAI処理ユニットと比較して優れた性能と効率を実現することが明確に示されています。NPUは、複雑なAIモデルを低レイテンシかつ高精度で処理できると共に、消費電力が大幅に少ないため、バッテリ駆動の組み込みデバイスへの適用に最適です。AIoT(AIとIoTを組み合わせた)コンスーマ製品およびスマートシティ・アプリケーションのケース・スタディは、Neural-ARTアクセラレータの実世界での利点を具体的に示しています。今後、STは、Neural-ARTアクセラレータの継続的な革新と改善に取り組んでいきます。将来のロードマップには、NPUファミリの性能、効率、汎用性をさらに高めることを目的としたエキサイティングな機能と機能強化が含まれています。デジタルNPUからインメモリ・コンピューティングへの移行は、エネルギー効率と計算性能の大幅な向上を約束し、次世代のニューラル・ネットワークの推論エンジンへの道を開きます。

参照資料

STは、ST Neural-ARTアクセラレータ™を使用して、開発者や企業が成功するために必要なサポートとリソースを提供しています。

エッジにおける人工知能

STがどのようにインテリジェンスをクラウドからエッジに移行しているかについて解説 **[STテクノロジー・ページ]**

ST Edge AI Suite

組み込みシステムにAI機能を統合するための包括的なツール・セット **[ST Developer Zone] [開発者ツール] [ケース・スタディ] [ST Community]**

ST Edge AI Model Zoo

組み込みアプリケーションにエッジAI機能を追加し、STデバイス上で動作するように最適化されたリファレンス・エッジAIモデルと、関連する展開スクリプト **[STM32 model zoo] [MLC model zoo on GitHub] [ISPU model zoo on GitHub]**

ST EdgeAI Developer Cloud

さまざまなSTデバイスでエッジAIモデルを簡単に最適化・ベンチマークできる無料のオンライン・プラットフォーム **[製品概要]**

STM32Cube.AI

ニューラル・ネットワークや機械学習モデルなど、事前に学習させたエッジAIアルゴリズムをSTM32用に最適化したCコードに自動変換できる無償のSTM32Cube拡張パッケージ **[製品概要]**

ST Edge AI Core

マイクロコントローラ、マイクロプロセッサ、MEMSセンサを含む複数のSTデバイス向けにエッジAIモデルを最適化およびコンパイルするコマンドライン・インタフェース(CLI)ツール **[製品概要]**

高速データログ

GUI(グラフィカル・ユーザー・インタフェース)、CLI(コマンド・ライン・インタフェース)、またはスマートフォンとのBluetoothを使用して、センサ・データセットの取得とラベリングを管理できる包括的なマルチセンサ・データ・キャプチャおよびビジュアライゼーション・ツールキット **[製品概要]**

ミックス
紙 | 責任ある森林管理を支えています
FSC® C002641
www.fsc.org

Order code: **WP2412NEURARTJ**

詳細については、STウェブサイトをご覧ください。www.st.com

■東京 TEL 03-5783-8200 ■大阪 TEL 06-6397-4130 ■名古屋 TEL 052-587-4547
STマイクロエレクトロニクス株式会社