GoAI を使用したエッジデバイスのためのフルスタック人工知能開発

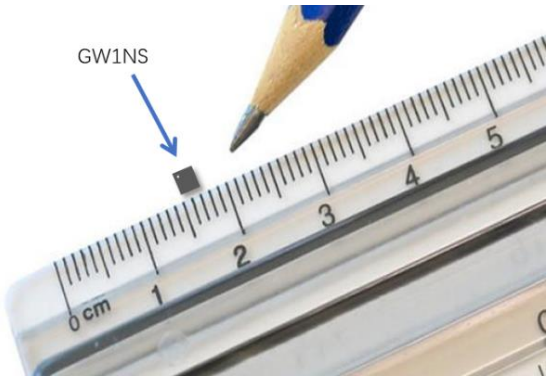
ホワイトペーパー

WP951-1.1J, 2020-09-21

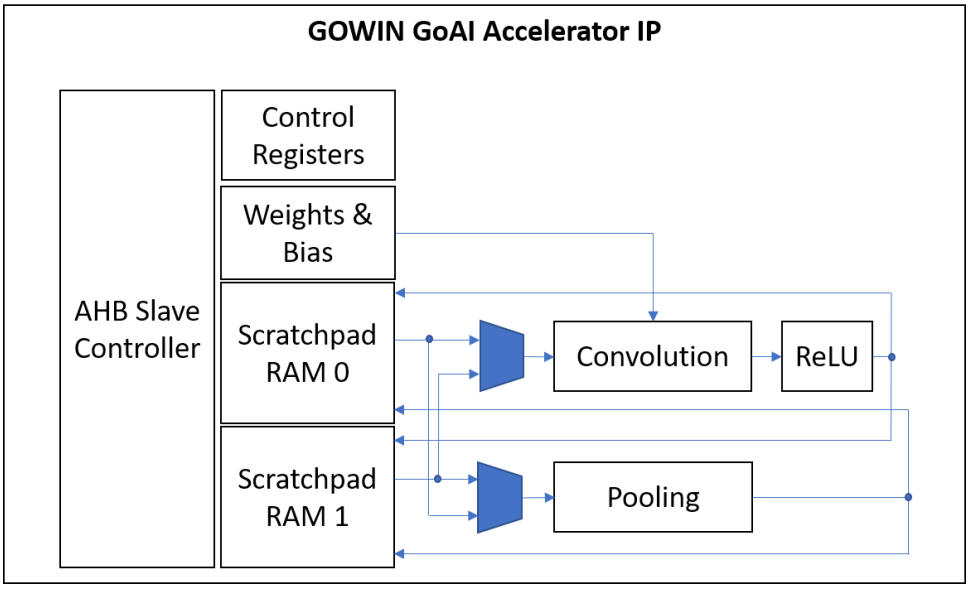
オートモーティブ、IoT、インダストリアルおよびコンシューマ向けの様々なタイプのアプリケーションでの、エッジデバイスへの役割の期待値は、急速に高まっています。これらのデバイスでは、エッジ推論が一般的な機能となり、接続されたノードでのローカルでの意思決定、遅延低減およびコスト削減を実現しています。

これらのソリューションは、コスト、消費電力、サイズ、および長期にわたる適合と統合のための柔軟性の顧客の更なる要求を満たすのに、苦労することがよくあります。さらに、ニューラルネットワークの重い計算要求は、多くの場合、標準マイクロコントローラの性能を超えています。これらのソリューションは、最新のテクノロジーの進歩を取り入れることを期待されながら、製品化のスケジュールを守ることも難しいです。

低集積度の FPGA を使用して、ニューラルネットワークのサイズに依存する、柔軟でスケーラブルなソリューションを提供することにより、コスト、消費電力、かつサイズに関する共通の顧客の制約を対処できます。GOWIN FPGA は、低消費電力と高性能のプロセス技術での、最小 3.24mm^2 のウェハーレベルの QFN および BGA パッケージング、および 1k から 55K LUT までの集積度のスケーラブルなデバイスを提供することにより、特に上記の難問に対処しています。

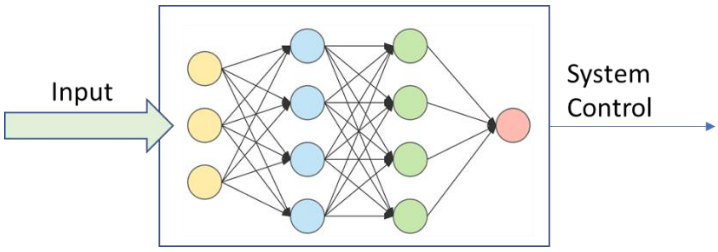


人工知能向けのエッジソリューション開発の性能と製品化までの時間を改善するために、GOWIN は自社の FPGA デバイスをターゲットとする「GoAI」という新しいアクセラレーション IP およびソリューションスイートを開発しました。GoAI ソリューションスイートは、GOWIN の AI アクセラレーション IP を既存の機械学習フレームワークに統合し、Cortex-M クラスのマイクロコントローラのみ使用する場合と比べて、性能を 78 倍以上に向上させました。

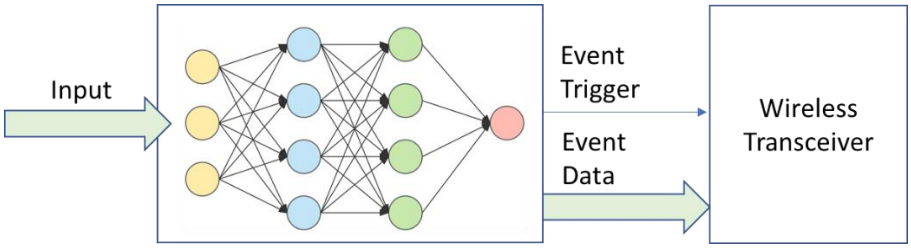

GOWIN GoAI Accelerator IP
AHB Slave Controller
Control Registers
Weights & Bias
Scratchpad RAM 0
Scratchpad RAM 1
Convolution
ReLU
Pooling

システムでのエッジ AI 使用

エッジの人工知能は、通常システム上、2 つの目的のいずれかで使用されます。　一つは、インターネットに接続されていないデバイスで推論を実行することです。これらのシステムは、機械学習を使用して入力情報を検知し、それを利用して、接続されているシステム出力を制御します。



二つ目の目的は、データをクラウドに送信してさらなる処理する前に、事前検知を実行することです。これはさまざまな都合で行うことができます。例えば、ワイヤレストランシーバーをオフにすることで電力を節約し、事前検知が発生したときにのみクラウド AI サービスにデータを送信することでコストを削減するなどです。



AI をエッジに実装

今日の人工知能は、畳み込みニューラルネットワークを中心とした機械学習テクノロジーを使用しています。これらのネットワークは本質的に、入力の重要な属性を特定するためにトレーニングされた係数または重みを持つ、多くのフィルターまたは「ニューロン」のセットです。これらの重みは、「トレーニング」と呼ばれるプロセスを通じて計算されます。このプロセスでは、入力データのセットが提供され、出力が既知が前提で、出力データを認識するために重みが更新されます。

畳み込みニューラルネットワークのトレーニングは、多くの場合、かなりの大量の計算パワーを必要とします。ただし、入力に関する特定の属性を推測するために使用される重みを生成するためだけで、通常はリアルタイムで実行する必要はありません。ネットワークをトレーニングすることにより、重みをネットワークにロードして、入力に関連する属性を検出できます。この推論は、多くの場合、トレーニングよりもかなり少ない計算パワーしか必要としません。

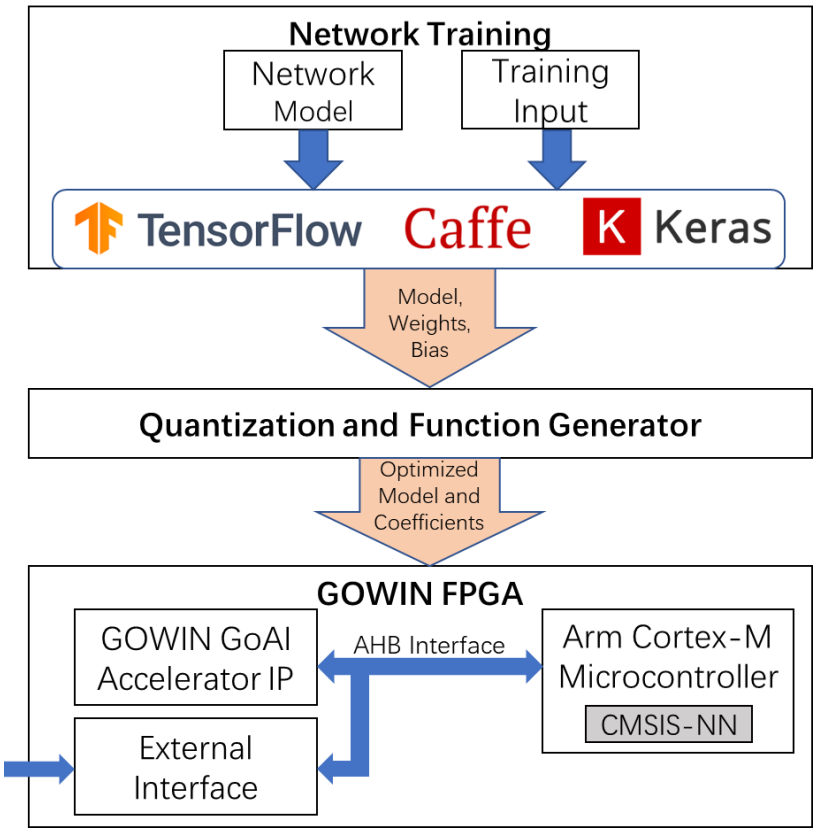
推論に必要な計算量はかなり少ないですが、たびたび、マイクロコントローラの性能を依然として上回ります。これは、マイクロコントローラがプロセッサのクロックサイクル

ごとの計算命令を処理する時間は 200MHz 以下であることが多く、小規模な機械学習ニューラルネットワークでさえリアルタイムでの検出ができないためです。さらに、AI に関連する多くのユース ケースでは、専用インターフェースとデータのバッファリングが必要です。例えば、フィルタリングは画像内の複数のピクセルに対して同時に実行されるため、カメラデータはフレームとして RAM に保存する必要があります。

エッジに焦点を当てた FPGA は、これらの問題に簡単に対処できます。ネットワークの並列およびパイプライン計算により、リアルタイム性が可能になり、数十 MHz でシステムをより効率的に運用できます。柔軟なインターフェースにより、FPGA はカメラ、マイクロフォン、生体認証センサー、その他の入力に簡単に接続できます。構成可能なメモリにより、中間データまたはデータ層のバッファリングと保持が可能になります。

FPGA はエッジ AI を可能にする優れた手段を提供しますが、開発とデプロイメントを容易にするには強力なソフトウェアスタックが必要です。ニューラルネットワークのモデリングソフトウェアは、いくつかのプロバイダー（Tensorflow、Caffe、Keras など）によって提供されています。これらのネットワークは、しばしば、ソフトウェアによるトレーニングとテストのために浮動小数点計算を使用して元々開発されています。これは、コストパフォーマンスに値するソリューションをエッジに実装しようとする時に問題を引き起こします。

その結果、マイクロコントローラ用の Tensorflow Lite や Arm CMSIS-NN などの一般的なデプロイメントツールは、最適化プロセスを使用して、トレーニング済みの重みデータを浮動小数点から 8 ビット固定小数点に切り捨て、量子化することで、エッジ重視のハードウェアに対して、リソースをより実用的にします。ただし、多くの場合、性能は依然として重要であるため、データ層の畳み込みとアキュムレーションのパイプラインを専用に処理するアクセラレータ設計が広く採用されています。これらのアクセラレータは、リアルタイム性能をさらに向上させるために ASIC または FPGA で設計できます。


Network Training
Network Model
Training Input
TensorFlow
Caffe
Keras
Model, Weights, Bias
Quantization and Function Generator
Optimized Model and Coefficients
GOWIN FPGA
GOWIN GoAI Accelerator IP
AHB Interface
Arm Cortex-M Microcontroller
CMSIS-NN
External Interface

システムの一例

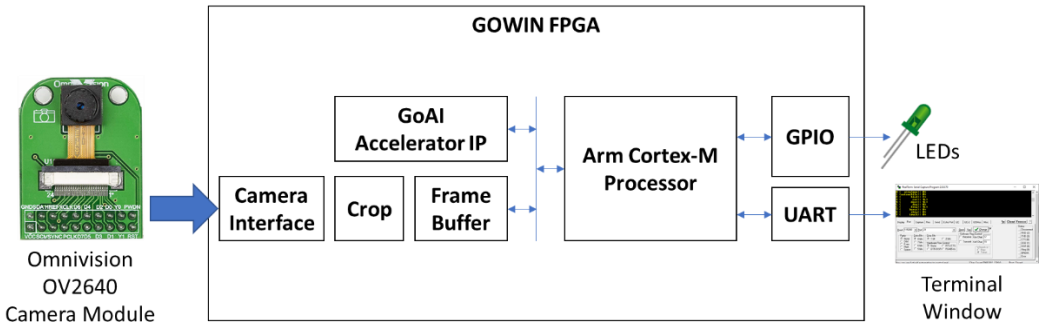
モデルトレーニングからハードウェア設計までの開発フロー全体を実行するには、GoAI プラットフォームを使用して CIFAR10 データセットで画像検出を実行しました。GoAI アクセラレータのパフォーマンスは、CMSIS-NN で同じネットワークを実行する Arm Cortex-M マイクロコントローラと比較されました。CIFAR10 データセットは、機械学習システムのさまざまなパフォーマンス属性を測定するために使用される 10 分類オブジェクトの一般的なデータの集まりです。

最初に、Caffe でシステム用のニューラルネットワークを訓練しました。この場合、テスト対象のネットワークとして、さまざまな数のフィルターを持つ 3 つの畳み込み層を使用しました。ネットワークがトレーニングされた後、重みとバイアスの係数が取得され、訓練されたネットワークはさまざまな入力を Caffe でテストされ、期待どおりに動作することを確認されました。

その後、スクリプト ユーティリティを使用して、重みとバイアス係数は切り捨て量子化され、ARM Cortex-M1 および M3 プロセッサで CMSIS-NN 関数として呼び出され使用されるように、ネットワークはコンパイルされました。

そして、最適化されたネットワークは、AHB バスに接続されたカメラインターフェースとフレームバッファを備えた ARM Cortex-M1 プロセッサに展開されました。ニューラルネットワークは、カメラからの 1 つの画像を処理するのに約 10 秒かかりました。

次に、GoAI アクセラレータは AHB バスに接続され、ネットワーク処理に使用されました。Cortex-M1 は、画像データを最初にアクセラレータに渡し、重みをロードしてバイアスをかけ、アクセラレータ設定を構成するために引き続き使用されました。ニューラルネットワークは GoAI アクセラレータを使用して処理するのに約 0.5 秒かかり、主に UART を介して送信される結果に関連する遅延が発生しました。



Arm Cortex-M3 プロセッサとアクセラレータでさらなる分析が行われました。Arm Cortex-M3 プロセッサを単独で使用する場合と GoAI アクセラレータを併用する場合の違いとして、約 78 倍のパフォーマンス向上することを示しています。

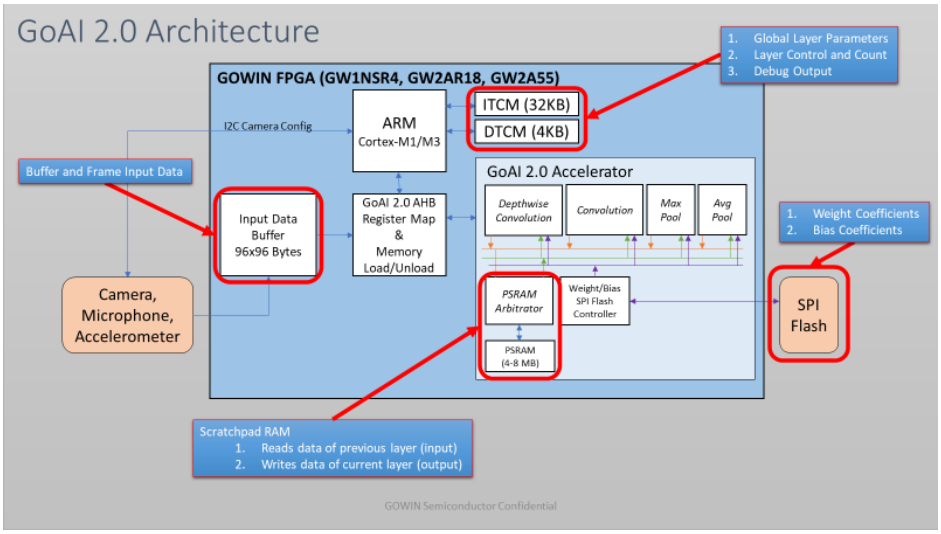
GoAI 2.0

GoAI 2.0 は以下に焦点を当てています。

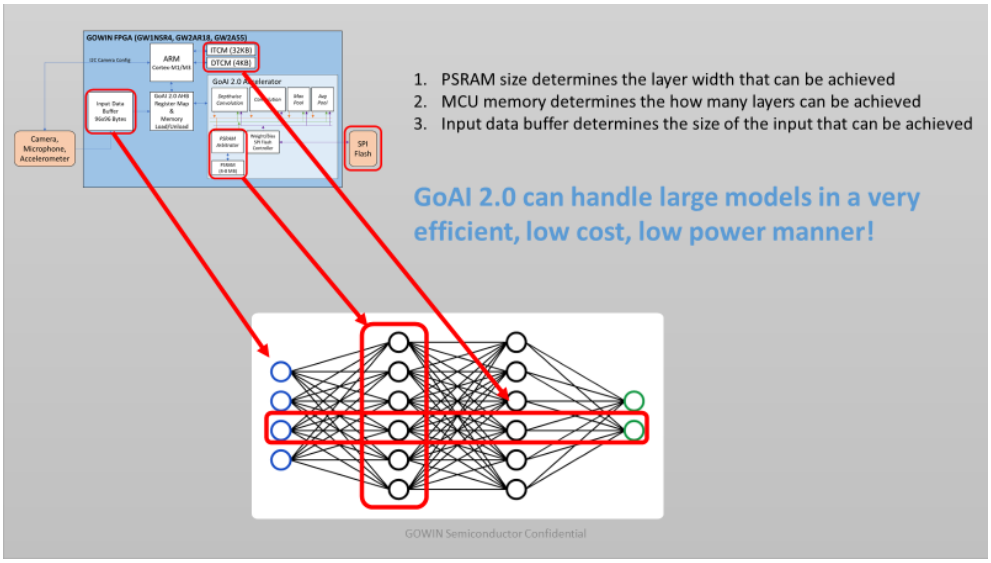
1. FPGA アクセラレータと TensorFlow および TensorFlow Lite の統合。
2. 6x6mm QFN パッケージの Cortex-M3 プロセッサ内蔵 GOWIN GW1NSR-4C uSoC FPGA をサポート。
3. ソフトウェアのコンパイルとデプロイメント SDK
4. 多数のレイヤと大きなレイヤ深度を持つ多様なモデルをサポートするための柔軟なアーキテクチャ

GoAI 2.0 プラットフォームは、標準的な TensorFlow 開発環境を使用して、任意のモデルのトレーニングとテストを可能にします。最終的にトレーニング済みモデルは、TFLiteConverter または TocoConverter を使用してモデルを解析し、量子化して *.tflite フラットバッファファイルにします。フラットバッファファイルは GoAI 2.0 SDK を使用して解析され、モデル係数、レイヤパラメータ、モデル関数を抽出します。

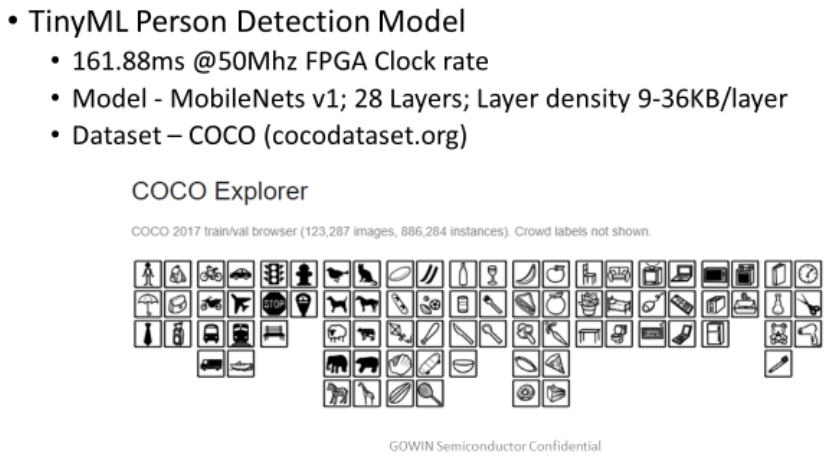
フラットバッファファイルから必要な情報を抽出した後、外部 SPI Flash メモリに係数を、Cortex-M3 のエンベデッド Flash に C コードを、GW1NSR-4C デバイスやその他の GOWIN FPGA にビットストリームをロードします。



GoAI 2.0 プラットフォームのアーキテクチャは、PSRAM が GW1NSR-4C に統合されており、レイヤ関連のパラメータを格納できるため、可能な限り最大のレイヤ深度と、可能な限り多くの畳み込み層およびプーリング層をサポートします。GW1NSR4 は 8MB の PSRAM を搭載しており、4MB の入力レイヤバッファと 4MB の出力レイヤバッファに分割されています。つまり、レイヤの入出力は最大 4MB のサイズになります。Cortex-M3 内の ITCM エンベデッド Flash は 32KB で、各レイヤの制御ループとフィルタパラメータを保存するだけで済みます。SPI Flash は、各レイヤのウェイトとバイアス係数を保存し、必要なモデルサイズに応じて調整できます。



GoAI 2.0 プラットフォームのテストは、Mobilenet v1.025 と COCO データセットを使用して行われました。Mobilenet は 28 層のかなり大規模な畳み込みニューラルネットワークです。GoAI 2.0 はこのモデルで 162ms の推論レイテンシを達成しました。



結論

合理的なコスト、消費電力、サイズ、および製品化までの時間内でエッジ AI を効率的に設計する時には、さまざまな課題が発生します。エッジでの人工知能は、接続されていないデバイスにとっても接続されているデバイスにとっても、ますます重要になっています。Edge AI ソリューションには、リアルタイム処理と一般的な機械学習モデル開発ソフトウェアへの統合のために、アクセラレータと完全なソフトウェア開発フローが必要です。GOWIN の GoAI アクセラレータとソフトウェアソリューション スタックは、性能やマーケット環境両方の制約に対処できる理想的なソリューションを提供しています。

テクニカル サポートとフィードバック

GOWIN セミコンダクターは、包括的な技術サポートをご提供しています。使用に関するご質問、ご意見については、直接弊社までお問い合わせください。

Web サイト: www.gowinsemi.com/ja

E-mail: support@gowinsemi.com

改訂履歴

日付	バージョン	説明
2019 年 9 月 16	1.0J	初版。
2020 年 9 月 30	1.1J	“GoAI 2.0”を追加。

免責事項

「GOWINSEMI®」、「LittleBee®」、「Arora」、及び GOWINSEMI のロゴは、当社により、中国、米国特許商標庁、及びその他の国において登録されています。商標又はサービスマークとして特定されたその他全ての文字やロゴは、www.gowinsemi.com において記載されているそれぞれの権利者に帰属しています。当社は、GOWINSEMI Terms and Conditions of Sale（GOWINSEMI 取引条件）に規定されている内容を除き、（明示的か又は黙示的かに拘わらず）いかなる保証もせず、また、知的財産権や材料の使用によりあなたのハードウェア、ソフトウェア、データ、又は財産が被った損害についても責任を負いません。本文書における全ての情報は、予備的情報として取り扱われなければなりません。当社は、事前の通知なく、いつでも本文書の内容を変更することができます。本文書を参照する何れの団体及び個人も、最新の文書やエラッタ（不具合情報）については、当社に問い合わせる必要があります。


GOWIN
PROGRAMMING FOR THE FUTURE