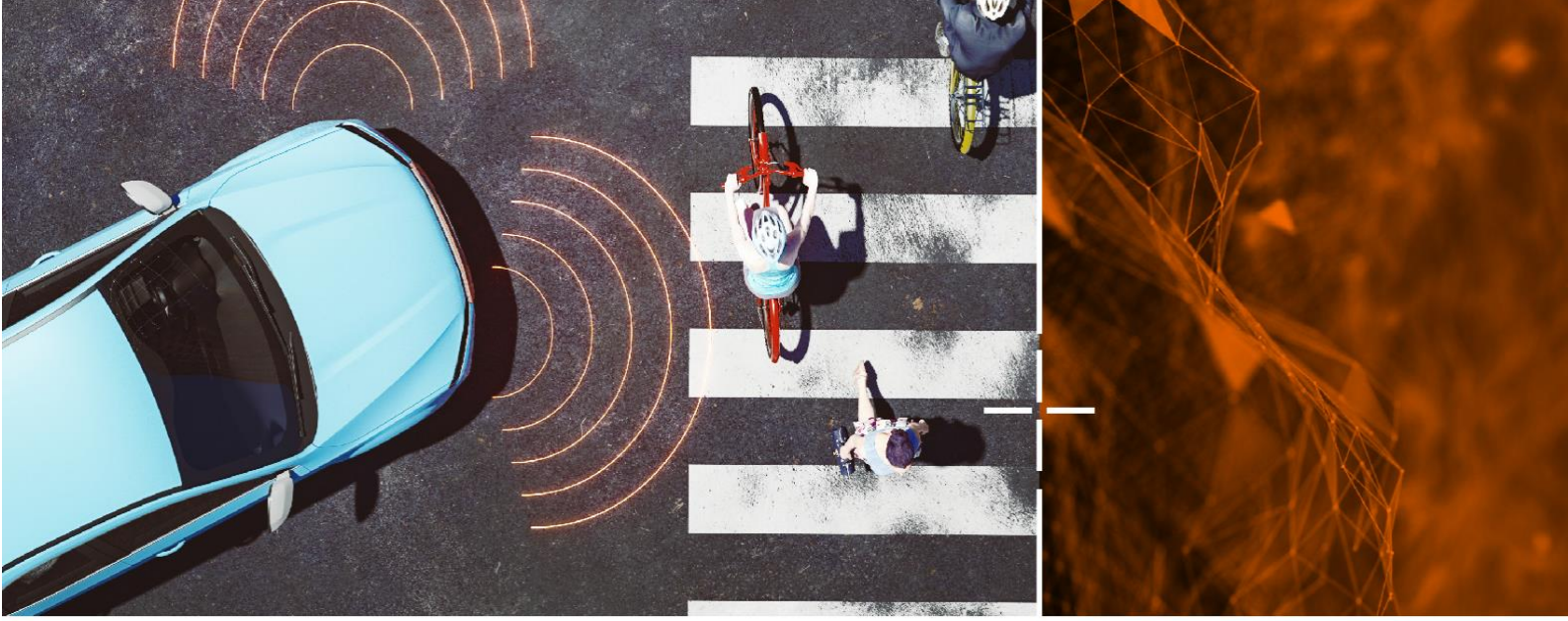


ケーススタディ

Plumerai が組み込みデバイスの人物検出機能に Arm Helium ベクトル拡張を採用

+ Plumerai
+ ソフトウェア
+ 従業員 25 人未満
+ 英国ロンドン
+ 2017 年設立

概要

現在、IoT 開発が極めて活発なのがイメージングの分野です。ビデオ付きインターホン、会議用モニター、住宅用セキュリティカメラ、スマートストアなど、人工知能（AI）や機械学習（ML）を利用したコスト効果の高いイメージングソリューションが、革新的な企業によって次々と開発されています。この分野の開発者にとって重要なのが精度と効率です。

「コンパクトなニューラルネットワークに対する Plumerai のアプローチは、垂直統合、およびすべての AI レイヤーをまとめて考えることです。つまりデータ、モデル、トレーニング、推論、ハードウェアを別々に扱うことはしません。効率を高めるには全体的に捉えることが重要です」

はじめに

英国ロンドンに本社を置く Plumerai は、小型の組み込みデバイスで、AI を利用した複雑なコンピュータービジョンタスクを効率的に実行することに特化しています。これには既知の顔、車、ペットを識別する人物検出機能が含まれます。同社のエンジニアはリアルタイム人物検出アプリを開発し、Arm Cortex-M85 コアをベースとする Renesas RA8D1 マイクロコントローラ（MCU）で実行するよう移植しました。ニューラルネットワークの高速処理には、このコアの Helium ベクトル拡張が活躍します。こうして同社は、極めて少ないシステムリソースで毎秒 13 フレームの高性能を達成しました。

課題

- 人物検出機能など、ニューラルネットワークを使用したコンピュータービジョンのタスクをリソースの少ない組み込みデバイスで処理するのは簡単ではありません。アプリには高い演算性能と小さなメモリフットプリントが求められます。
- マイクロコントローラはもともと、SIMD 命令による並列処理や高速処理に対応していません。この欠点を補うために Arm Helium ベクトル拡張（MVE: Arm M-Profile ベクトル拡張）が採用されました。
- 全データをデバイスに保存し、クラウド接続に依存しなければ、ユーザーのプライバシー保護とソリューションのセキュリティが確保されます。

ソリューション

Plumerai は Arm Cortex-M85 の Helium ベクトル拡張を利用し、人物検出ニューラルネットワークを高速化しました。

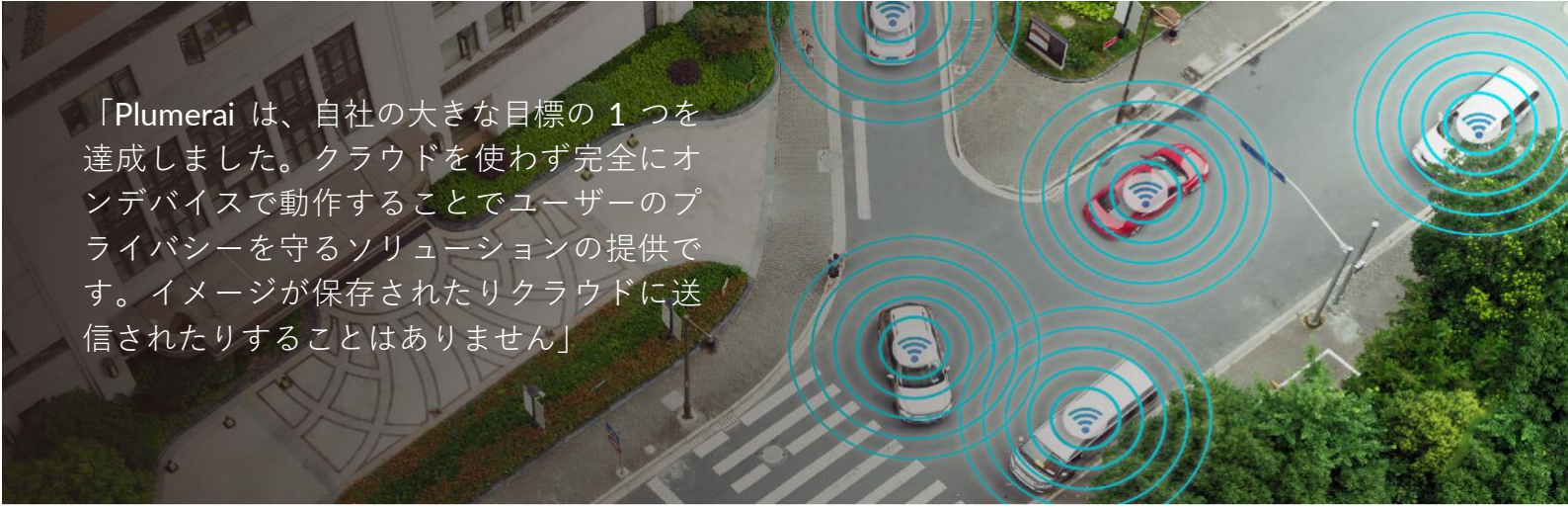
コンパクトニューラルネットワークに対する Plumerai のアプローチは、垂直統合、およびすべての AI レイヤーをまとめて考えることです。つまりデータ、モデル、トレーニング、推論、ハードウェアを別々に扱うことはしません。効率を高めるには全体的に捉えることが重要です。

モデルアーキテクチャはプロセス全体の一部にすぎません。このアプローチでは、それだけに力を注ぐのではなく、各要素とデータの絡み合いを考えます。小さなニューラルネットワークにとってはデータがすべてであり、トレーニングデータの収集、選別、適切なラベル付けが重要となります。

Plumerai は、応用範囲が広く、エコシステムが充実していることから Arm アーキテクチャを選択しました。Arm Cortex-M MCU でソフトウェアを実行し、毎秒 2〜5 フレームに達するイメージキャプチャ性能を確保しました。

そして Cortex-M85 Helium 拡張を備えた高性能の RA8x1 MCU が Renesas から発表されたとき、強い関心を持ちました。

Arm Helium は Cortex-M クラスのプロセッサーに対応する拡張機能であり、AI/ML ワークロードを小型で高電力効率のデバイスで実行する能力を高めます。Helium は、最適化されたハードウェアとソフトウェアによって Cortex-M プロセッサー上でニューラルネットワークモデルの実行速度を高めるため、スマートセンサー、IoT デバイス、ウェアラブルなど多様な用途に適しています。

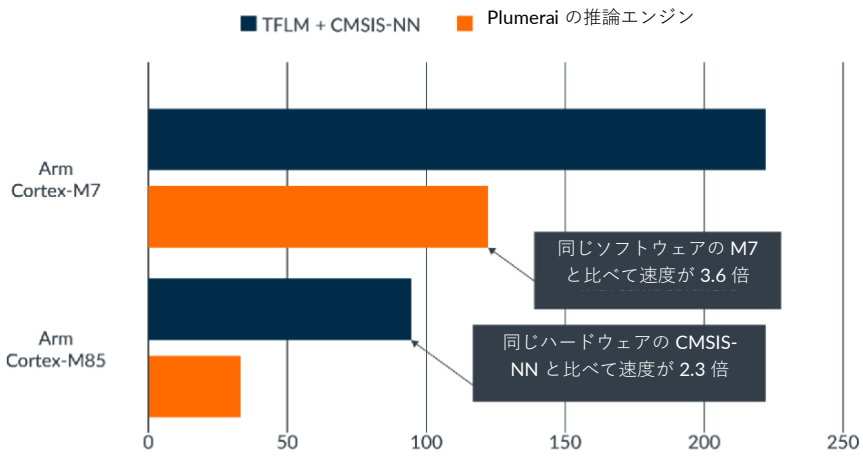


「Plumerai は、自社の大きな目標の 1 つを達成しました。クラウドを使わず完全にオンデバイスで動作することでユーザーのプライバシーを守るソリューションの提供です。イメージが保存されたりクラウドに送信されたりすることはありません」

Plumerai は、Helium の広いベクトルレジスター、および EDP（enhanced dot product）などの新しい SIMD 命令を駆使し、畳み込みや行列乗算などの重要なニューラルネットワークレイヤーを高速化しました。

Plumerai の最適化された推論エンジンも性能向上に大きな役割を果たします。Helium を使用したマイクロコントローラの速度は TensorFlow Lite の 3.5 倍となりました。Helium による高速化で人物検出の速度は 4 倍になり、わずか 480 MHz で動作する Cortex-M85 ベースの RA8D1 MCU で 13 FPS もの性能を実現しました。

Cortex-M85 ベンチマークでの人物検出



推論あたりのクロックサイクル数（単位は百万、低いほど良い）

これはイメージング速度だけでなく精度も高めるため、人数計測など新しい用途の可能性もあります。システム消費電力全体の引き下げにも有効です。フレームレートが高いということは、フレームを短時間でキャプチャし、分析できるということです。フレーム内に人が検出されなければシステムはすぐにスリープに入れます。

Plumerai はこれをわずか 300 KB RAM で実現し、Cortex-M85 ベースの Renesas RA8D1 MCU 評価ボードに実装しました。極めて複雑な AI ビジョンタスクにもかかわらず、実行可能バイナリサイズは全部でわずか 1.5 MB です。2MB のフラッシュメモリ、1MB の SRAM、16 ビットのカメラインタフェースを持つ RA8D1 MCU が、外付けメモリや他のコンポーネントなしで包括的な人物検出ソリューションを実現しました。人物検出のクオリティは高く、オクルージョン、各種のポーズ、照明の状態など、難しい条件にも対応します。Plumerai は、自社の大きな目標の 1 つを達成しました。クラウドを使わず完全にオンデバイスで動作することでユーザーのプライバシーを守るソリューションの提供です。イメージが保存されたりクラウドに送信されたりすることはありません。

Plumerai は独自の最適化された推論エンジンとフレームワークも構築しました。これは ML Commons により世界最速と記録されています。

まとめ

Plumerai は、Arm Helium ベクトル拡張を利用し、極めてリソースの少ない組み込みデバイスで高性能のコンピュータービジョン処理を可能にしました。この実装は Helium の力を実証するとともに、組み込み AI ワークロードに携わる他の開発者のモデルともなっています。

参考リンク

- Plumerai の人物検出ソリューション
- Arm Cortex-M85 と Arm Helium テクノロジー
- Plumerai の人物検出ソリューションを動画で紹介 – Arm Tech Talks Renesas RA8D1 マイクロコントローラ