CASE STUDY

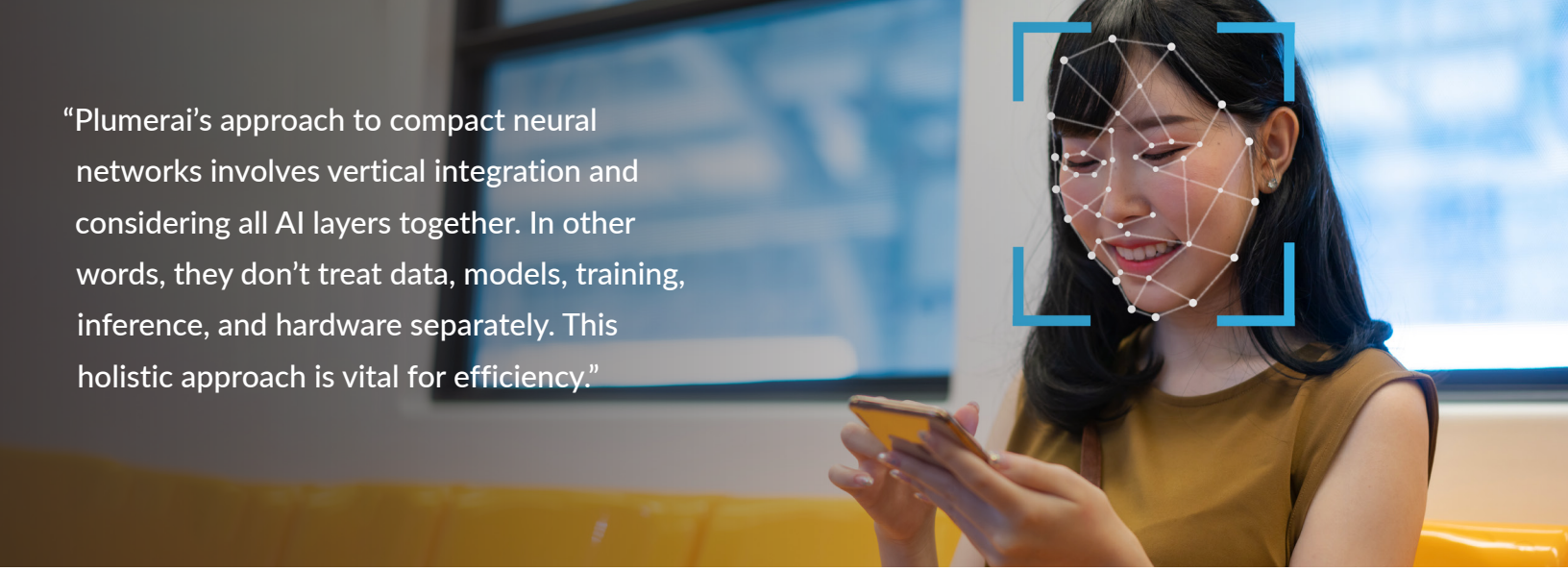
Plumerai's People Detection on Embedded Devices with Arm Helium Vector Extensions

Plumerai

- Plumerai
- Software
- <25 employees
- London, UK
- Founded 2017

OVERVIEW

One of the hottest areas of IoT development today is in imaging. Whether it's video doorbells, conference-room monitoring, home security cameras or smart retail applications, innovative companies are developing cost-effective imaging solutions that leverage artificial intelligence (AI) and machine learning (ML). Increasingly important for developers in this area is accuracy and efficiency.



"Plumerai's approach to compact neural networks involves vertical integration and considering all AI layers together. In other words, they don't treat data, models, training, inference, and hardware separately. This holistic approach is vital for efficiency."

Introduction

Plumerai, headquartered in London, specializes in enabling complex AI-assisted computer vision tasks efficiently on small, embedded devices. These tasks include the detection of people, including identifying familiar faces, vehicles, and pets. Its engineers developed a real-time people detection application and ported it to run on the Renesas RA8D1 microcontroller (MCU) based on the Arm Cortex-M85 core, making use of the Helium vector extensions to accelerate the neural network. This enabled them to achieve high performance of 13 frames per second with minimal system resources.

Challenges

- Running neural network-based computer vision tasks like people detection on resource-constrained embedded devices is challenging. Such applications require high compute performance and low memory footprint.
- Microcontrollers traditionally lack support for SIMD instructions to exploit parallelism and accelerate performance. Arm Helium vector extensions (also called Arm M-Profile Vector Extensions) were introduced to address this gap.
- Not relying on a cloud connection by keeping all data on-device to ensure user privacy and enhance solution security.

Solution

Plumerai leveraged Helium vector extensions on the Arm Cortex-M85 to accelerate their people-detection neural networks.

Plumerai's approach to compact neural networks involves vertical integration and considering all AI layers together. In other words, they don't treat data, models, training, inference, and hardware separately. This holistic approach is vital for efficiency.

Their approach doesn't just focus on the model architecture; that's just a fraction of the entire process. They consider how components are intricately tied to data. Data is crucial for tiny neural networks and gathering, curating and correctly labeling training data is vital.

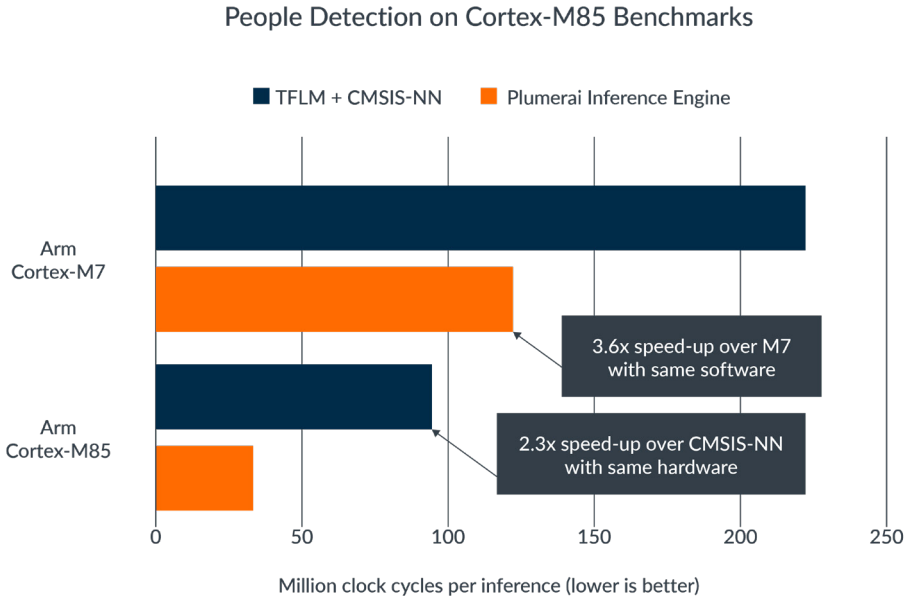
Plumerai selected the Arm architecture because of its extensive reach and ecosystem. They've run software on Arm Cortex-M MCUs, achieving solid image-capture performance reaching 2 to 5 frames per second. They were intrigued when the high performance RA8x1 MCUs with Cortex-M85 Helium extensions were announced.

Arm Helium is a technology extension for Cortex-M-class processors that provides enhanced capabilities for executing AI and ML workloads on small, power-efficient devices. Helium includes hardware and software optimizations that help achieve faster execution of neural network models on Cortex-M processors, making them suitable for various applications, including smart sensors, IoT devices, wearables, and more.

"Plumerai was also able to deliver on one of the company's key user objectives: Delivering a privacy-friendly solution that runs completely on-device without cloud offload. No images are stored or transmitted to the cloud."

They used Helium's wider vector registers and new SIMD instructions like EDP (enhanced dot product) to speed up key neural network layers like convolutions and matrix multiplications.

Plumerai's optimized inference engine also played a key role in maximizing performance. It was measured to be 3.5x faster than TensorFlow Lite for Microcontrollers with Helium. Plumerai achieved a 4x speedup of people detection through Helium acceleration, boosting performance to 13 FPS on the Cortex-M85 based RA8D1 MCU running at just 480 MHz.



This not only increases imaging speed but also enhances accuracy, which, in turn, opens up new applications like people counting. It also can be leveraged to lower overall system power: Higher frame rates mean capturing and analyzing frames faster, so if no person is detected in the frame, the system can go to sleep sooner.

Plumerai was able to implement this on a Renesas evaluation board with Cortex-M85 based RA8D1 MCU, using just 300 KB RAM. Their entire executable binary size is only 1.5 MB for an extraordinarily complex AI-vision task. The RA8D1 MCUs with 2MB of flash and 1MB of SRAM memory on-chip and a 16bit camera interface, enabled the comprehensive people detection solution without need for any external memory or other components. The quality of people detection is high, handling difficult cases like occlusion, different poses and difficult lighting situations. Plumerai was also able to deliver on one of the company's key user objectives: Delivering a privacy-friendly solution that runs completely on-device without cloud offload. No images are stored or transmitted to the cloud.

Additionally, the company also built its own optimized inference engine and framework that has been benchmarked by ML Commons as the fastest in the world.

Conclusion

Arm Helium vector extensions have enabled Plumerai to unlock high-performance computer vision applications on extremely resource-constrained embedded devices. Their implementation demonstrates the capabilities of Helium and serves as a model for other developers working on embedded AI workloads.

Useful Links

- Plumerai People Detection Solutions
- Arm Cortex-M85 and Arm Helium Technology
- Watch Plumerai's People Detection Solutions in Action – Arm Tech Talks
- Renesas RA8D1 Microcontrollers

arm