Increasing Intelligence at the Edge With Embedded Processors

Texas Instruments

Alec May
Systems Manager
Jacinto™ High-Performance Computing Processors

At a glance

Introduction

The accessibility and user-friendliness of widely available cloud-based AI solutions has made it easier for almost anyone to engage with models and tools designed for AI.

Not all AI innovations are happening in the cloud, however. With technological advancements in embedded processor design, AI capabilities are now in consumer products such as laptops and cellphones, as well as other battery-powered applications: video doorbells, vision processing in automotive systems, and motors for energy infrastructure and industrial systems.

Edge AI – the ability to run AI models locally, near the source of the data – is enhancing the responsiveness, efficiency, reliability and security of these products. The embedded processors making the cloud-to-edge transformation possible integrate components such as specialized cores for digital signal processing (DSP) and are supported by easy-to-use GUI-based tools that minimize the time and expertise required to bring AI to the edge.

This white paper explores the evolution and benefits of edge AI, as well as the advancements in hardware and software that are enabling it.

What is AI?

When most people today think of AI, they will often imagine text and image generators. But even the simplest of algorithms is technically an example of AI in the literal sense.

The broadness of AI and its multiple use cases have led to several subdomains, including machine learning and deep learning, as shown in Figure 1.

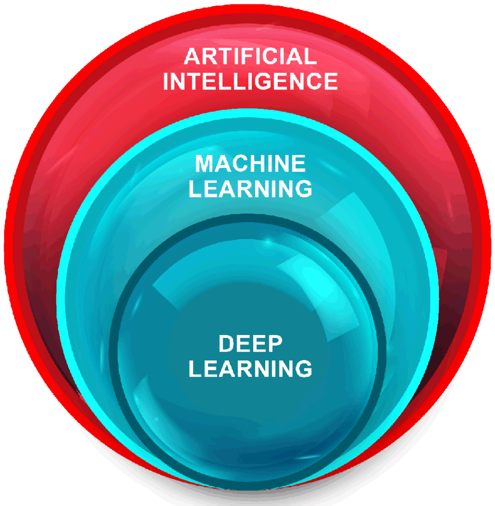

Figure 1. *The relationship between different AI subdomains.*

The majority of AI used for embedded applications is machine learning, the subdomain where machines and algorithms "learn" how to solve a problem from data; for example, a vehicle recognizing a pedestrian vs. an obstacle by analyzing image data for common patterns. A machine learning model learns from training data, which may be labeled with ground truth information (i.e. verified, accurate data) to better identify which patterns to learn from. This training process enables machine learning models to discern patterns in the data, which they can use to make future inferences.

Within the field of machine learning, deep learning has become one of its most popular implementations given its ability to solve highly complex problems accurately, although doing so requires plenty of computing resources. Deep learning uses multilayer neural networks, which are data models inspired by the neurons in the human brain. Neural networks enable developers to solve problems where the patterns are too complex to discern or write custom rules for.

What are the benefits of edge AI?

AI and its subdomains can typically perform processing either in the cloud or on local servers. Cloud-based AI has historically been more common, since the computing power needed to perform impactful AI was not easily achievable outside of large servers. Edge AI has grown in popularity, however, with the increasing computing power and power efficiency of embedded processors.

Figure 2 shows how edge AI and cloud AI differ in regard to how they receive and process data and interact with cloud-based resources.

Edge AI often uses cloud or desktop resources for model training during development. After deploying the model to the embedded device, it becomes possible to make model inferences and decisions on new data independently and locally.

Until recently, most meaningful examples of AI required processing capabilities beyond what average consumer electronics could provide. This meant that machine learning models were often trained and implemented on cloud-based resources. While cloud-based implementations provide convenience by minimizing hardware investment, they have also limited the adoption of AI. Cloud-based AI implementations will not work in any application without cloud access – in other words, a network connection. In addition, edge AI can also improve security, safety and responsiveness compared to cloud-based AI.

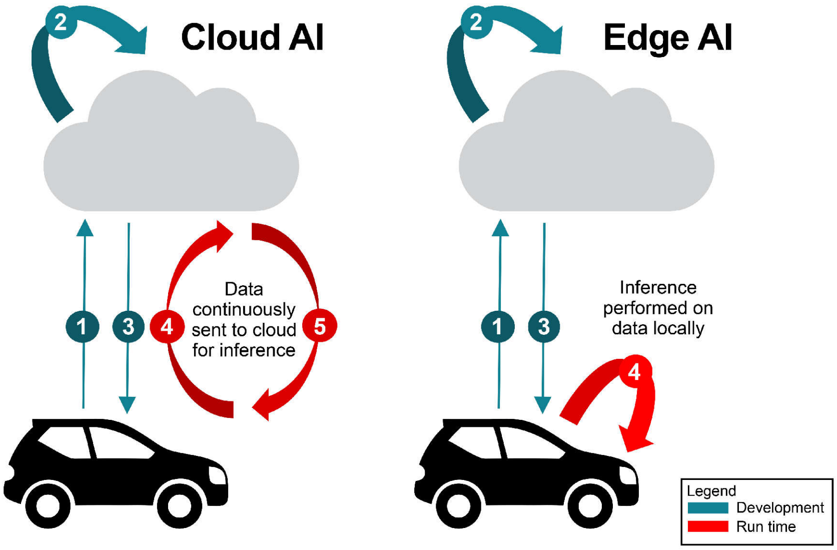
Cloud AI
2
1
3
4
Data continuously sent to cloud for inference
5
Edge AI
2
1
3
Inference performed on data locally
4
Legend
Development
Run time

Figure 2. *Comparison of cloud-based AI and edge AI.*

With advances in semiconductors as well as improvements to AI toolchains, the implementation of AI solutions directly into embedded processors and microcontrollers (MCUs) brings AI to the edge. Bringing AI to the edge means that computation and AI inference occur closer to the sensors providing data, which is important considering the increasing amount of sensor data gathered by electronics. Growing data volumes make cloud-only resources less practical, since the transmission of higher volumes of data to and from the cloud can be costly, complex, and represent a single point of failure.

Running AI models at the network edge typically reduces the latency for inference and decision-making based on sensor data; for example, a camera sensor in a vehicle for collision detection. With edge AI capabilities, a vehicle can make inferences faster, responding to stimuli in real time, without waiting for the inference from the cloud. This local inference is turned into action through physical AI (as shown in **Figure 3**), a term that refers to systems that sense, interpret and respond with real-world actions, such as a robot moving a box on a factory floor or a vehicle automatically applying the brakes.

Figure 3. *Simplified comparison of edge AI and physical AI in a humanoid robot*

These systems can use edge AI to blend advanced perception with mechanical actuation, enabling machines to operate safely and collaboratively alongside humans.

Edge AI has several other advantages compared to cloud-based AI, including reduced reliance on network connectivity. Edge AI works in applications where access to the cloud is not possible and minimizes potential downtime caused by network outages. Also, since cloud-based AI requires network connectivity, there can be recurring service fees for access, which can be a challenging business model when designing consumer products.

How AI is moving to the edge

The processing and power-consumption constraints of embedded processors, as well as the high level of in-house programming expertise and resources, have limited the broad adoption of edge AI. Embedded devices capable of meeting the performance requirements of AI calculations were often too large, consumed too much power, and generated too much heat.

Specialized hardware solutions emerged, enabling better acceleration for the computational operations needed to enable edge AI, but several trade-offs limited their wider adoption in edge applications. Dedicated hardware solutions such as graphics processing units (GPUs), field-programmable gate arrays (FPGAs) and application-specific integrated circuits (ASICs) have achieved impressive performance. These solutions are typically limited, however, by either high power consumption (especially in the case of GPUs and FPGAs) or limited flexibility (in the case of ASICs).

Integrated neural processing units (NPUs) have emerged as a solution for bringing AI capabilities directly to embedded systems. Unlike general-purpose processors, NPUs are purpose-built to execute the matrix multiplications, convolution operations and activation functions that form the backbone of modern neural networks. By offloading computationally intensive tasks from the main CPU, NPUs dramatically improve inference speed while reducing power consumption, two essential requirements for edge deployment.

To understand the impact of these components, let's look at two NPUs from Texas Instruments designed for different segments of the edge AI market: the C7™ NPU for high-performance applications and the TinyEngine™ NPU for low-power, scalable devices.

The C7 NPU

The C7 NPU is a high-performance, power-efficient AI accelerator integrated into the TDA54-Q1 and TDA4VE-Q1 systems-on-a-chip (SoCs). TI microprocessors and SoCs have included versions of the C7 NPU for several generations to address the computational needs of vision applications. This NPU comes from TI's long history of DSPs, which allows it to efficiently power AI solutions without sacrificing cost or power.

C7 NPUs also enable TI processors to handle multiple concurrent AI workloads, which is an important capability for systems that need to process data from cameras, radar, lidar and other sensors simultaneously.

The TDA54-Q1 uses the C7 NPU to enable edge AI in advanced driver assistance systems, infotainment and robotics.

The TinyEngine NPU

The TinyEngine NPU (as shown in the simplified block diagram in Figure 4) is a dedicated hardware accelerator for MCUs that optimizes deep learning inference operations to reduce latency and power consumption when processing AI workloads in resource-constrained devices, including battery-powered products.

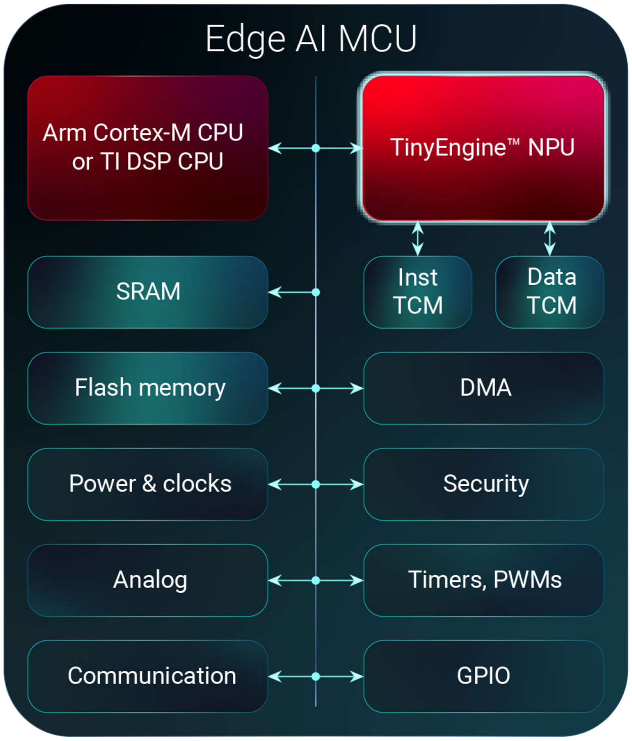

Figure 4. Simplified block diagram of a TI edge AI MCU with an integrated TinyEngine NPU

This NPU executes machine learning algorithms in parallel to the primary CPU running application code. MCUs with a TinyEngine NPU can run models with up to 90 times lower latency and 120 times lower energy utilization per inference than similar MCUs without an accelerator.

The TMS320F28P550SJ C2000™ MCU uses the TinyEngine NPU for motor bearing and solar arc fault detection, leaving the main CPU to handle real-time motor control. The AM13E23019 combines the TinyEngine NPU with an advanced real-time control architecture (in as many as four motors) for adaptive control and predictive maintenance in appliances, robotics and industrial systems. The MSPM0G5187

Arm® Cortex®-M0+ based MCU uses its dedicated TinyEngine NPU to execute deep neural network models independently from the main CPU, enabling edge AI capabilities in wearable health monitors, appliances and industrial systems for predictive motor maintenance.

Edge AI software innovations

Along with the hardware advancements for efficient AI computation in embedded devices, open-source communities and semiconductor manufacturers are also making it easier to test and deploy AI models with minimal programming expertise. Making AI more user-friendly (and in some cases GUI-based) helps reduce the need to invest in additional resources or training.

For designers who have more familiarity with AI models, open-source tools such as PyTorch and TensorFlow can train a model architecture for their custom dataset, and can export the model into an embedded-friendly format such as ONNX or LiteRT, formerly known as TensorFlow Lite. The model then runs with equivalent open-source runtime software on the device.

These open-source tools aid edge AI development by abstracting away specifics of the embedded platform, offering a consistent interface that allows access to hardware acceleration backends (also known as delegates). These backends can expose further configurations to give designers more control over model delegation to the hardware accelerator.

TI's **CCStudio™ Edge AI Studio** is a collection of web-based tools that simplifies and accelerates the development of edge AI applications on TI embedded devices using remote TI hardware and a GUI. The tools include a model composer, model analyzer, model selection tool and model maker, which can help designers quickly evaluate models and their performance without physically connecting to an evaluation board.

The scalability of edge AI

When developing a product using an embedded microcontroller or microprocessor, it's always important to consider how the product may evolve and scale over time. Engineers don't want to spend months developing a solution on one microprocessor and then have to start from scratch when they update their product to a higher-performance processor.

Semiconductor manufacturers that create these embedded devices need to develop portfolios with scalability in terms of features, performance and cost. This approach helps ensure that there is a seamless migration strategy between their various embedded processors for AI, in order to make it as simple as possible for developers to reuse their work across different devices.

Edge AI is no exception. For example, a designer making a home robot may want to produce both a high-end version with three cameras for surround vision and an entry-level version that only has a single front camera. A scalable portfolio of edge AI-accelerated devices enables the porting of software from the high-end model to the entry-level model, minimizing the amount of resources needed to produce both products. Scalability also allows developers to transfer their R&D investment from one platform to the next as their product evolves.

Conclusion

Although edge AI is still relatively new, its potential to reshape our daily lives is coming into focus, especially its ability to bring more responsiveness and performance to almost any application. With advancements in low-power, cost-effective embedded processors and intuitive software and model training tools, the barrier to entry for designers of any experience level has never been lower. We can expect this to continue with each successive generation of edge AI devices and the crucial components (for example, semiconductors for sensing, power delivery and connectivity) that manage the operation and data collection within the electronics we interact with and rely on.

Additional resources

- Explore ***TI's edge AI processing portfolio and design resources***.
- Learn about TI's edge AI-accelerated MCUs in the following technical articles
  - ***How edge AI-accelerated Arm® Cortex®-M0+ MCUs bring more brain power to electronics***
  - ***Achieving edge AI-enabled motor control in industrial automation and home appliance designs***
- Read about the TinyEngine NPU and its benefits for edge AI-accelerated embedded designs in the product overview, ***TI's TinyEngine™ NPU unlocks edge AI acceleration in more embedded systems.***

Important Notice: The products and services of Texas Instruments Incorporated and its subsidiaries described herein are sold subject to TI's standard terms and conditions of sale. Customers are advised to obtain the most current and complete information about TI products and services before placing orders. TI assumes no liability for applications assistance, customer's applications or product designs, software performance, or infringement of patents. The publication of information regarding any other company's products or services does not constitute TI's approval, warranty or endorsement thereof.

IMPORTANT NOTICE AND DISCLAIMER

TI PROVIDES TECHNICAL AND RELIABILITY DATA (INCLUDING DATASHEETS), DESIGN RESOURCES (INCLUDING REFERENCE DESIGNS), APPLICATION OR OTHER DESIGN ADVICE, WEB TOOLS, SAFETY INFORMATION, AND OTHER RESOURCES "AS IS" AND WITH ALL FAULTS, AND DISCLAIMS ALL WARRANTIES, EXPRESS AND IMPLIED, INCLUDING WITHOUT LIMITATION ANY IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT OF THIRD PARTY INTELLECTUAL PROPERTY RIGHTS.

These resources are intended for skilled developers designing with TI products. You are solely responsible for (1) selecting the appropriate TI products for your application, (2) designing, validating and testing your application, and (3) ensuring your application meets applicable standards, and any other safety, security, regulatory or other requirements.

These resources are subject to change without notice. TI grants you permission to use these resources only for development of an application that uses the TI products described in the resource. Other reproduction and display of these resources is prohibited. No license is granted to any other TI intellectual property right or to any third party intellectual property right. TI disclaims responsibility for, and you fully indemnify TI and its representatives against any claims, damages, costs, losses, and liabilities arising out of your use of these resources.

TI's products are provided subject to TI's Terms of Sale, TI's General Quality Guidelines, or other applicable terms available either on ti.com or provided in conjunction with such TI products. TI's provision of these resources does not expand or otherwise alter TI's applicable warranties or warranty disclaimers for TI products. Unless TI explicitly designates a product as custom or customer-specified, TI products are standard, catalog, general purpose devices.

TI objects to and rejects any additional or different terms you may propose.