Increasing Intelligence at the Edge With Embedded Processors

Texas Instruments

Alec May
Product Marketing Engineer
Embedded Processors

At a glance

Introduction

AI has moved from a niche technology to something people interact with daily, growing beyond just the engineering and technology sectors. This trend has led companies in nearly every industry to consider how they could leverage AI to increase efficiency, reduce costs, and increase their product's capabilities. 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 making their way into consumer products such as laptops and cellphones as well as other electronics: battery-powered applications like 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 electronics. The embedded processors making this 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.

In this white paper, I will discuss the evolution and benefits of edge AI, as well as the advancements in hardware and software that are enabling it.

What is AI?

AI is a machine's ability to exhibit some sort of intelligence or reasoning. When most people today think of AI, they will often imagine text and image generators or virtual opponents in video games. 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.

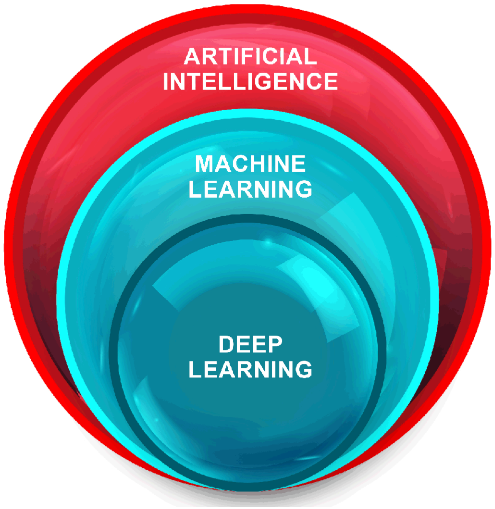

Figure 1. *The relationship between different AI subdomains.*

The majority of AI used for embedded applications is machine learning, the subdomain where machines and algorithms “learn” how to solve a problem; for example, a vehicle recognizing a pedestrian vs. an obstacle by analyzing image data for common patterns. A machine learning model can also learn by receiving large amounts of training data, or data that is already labeled. 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. It enables product designers to create solutions that are able to recognize patterns that they could not discern themselves.

What is edge AI?

AI and its subdomains can typically perform processing either in the cloud or on local hardware. 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.

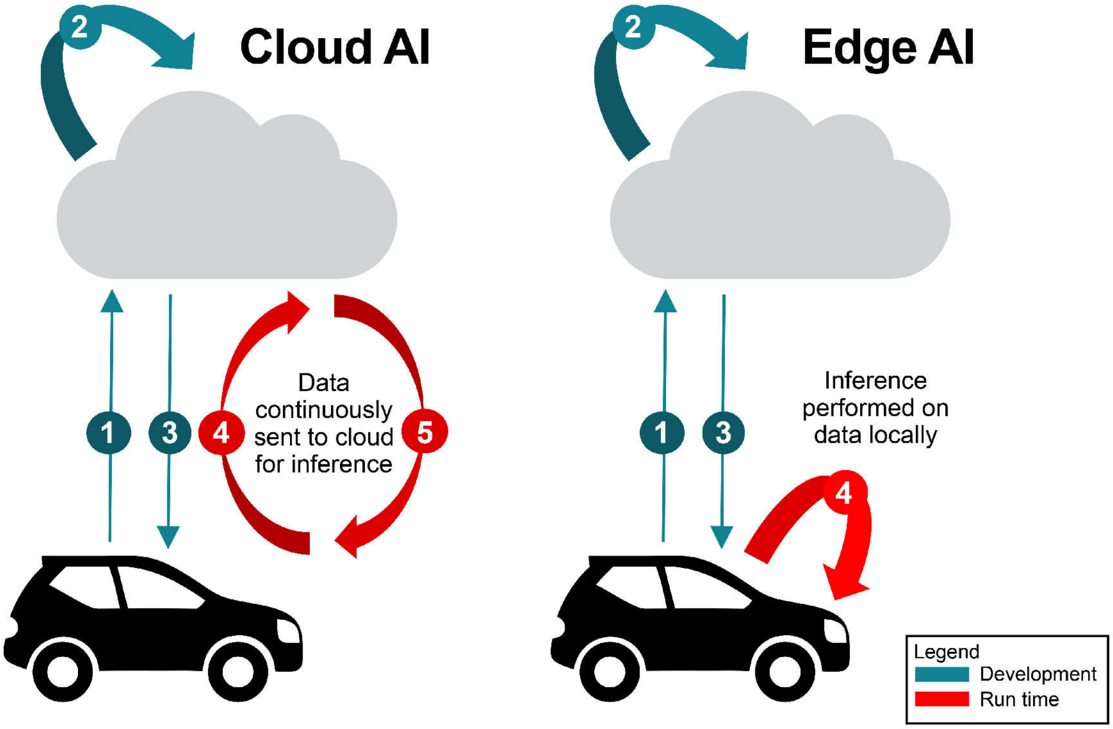
Cloud AI
2
1
3
4
Data continuously sent to cloud for inference
5
Edge AI
2
1
3
Inference performed on data locally
4
Legend
🟦 Development
🟥 Run time

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

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 cannot be used 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.

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 are performed closer to the sensors providing data, which is important considering that the amount of sensor data gathered by electronics continues to increase. 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.

Edge AI has several other advantages compared to cloud-based AI, including reduced reliance on network connectivity. Edge AI can be used 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

Let's look at how advancements in embedded processors and user-friendly, web-based software is helping more designers enable edge AI capabilities in their designs.

Edge AI hardware innovations: In the past, the processing and power consumption constraints of embedded processors, as well as the high level of in-house programming expertise and resources, 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.

In recent years, specialized hardware solutions have emerged, enabling better acceleration for the computational operations needed to enable edge AI. These hardware solutions had several trade-offs that limited their wider adoption in edge applications, however. Dedicated hardware solutions such as graphics processing units (GPUs), field-programmable gate arrays (FPGAs) and application-specific integrated circuits (ASICs) have been able to achieve impressive results, but are typically limited by their power consumption (especially in the case of GPUs and FPGAs) or their lack of flexibility (in the case of ASICs).

Newer embedded processors with integrated edge AI capabilities address these power limitations and cost considerations. Many of these devices feature integrated components like a neural processing unit (NPU) and/or digital signal processors (DSP).

Figure 3 shows an example of a TI C2000™ MCU with an integrated NPU for motor and solar arc fault detection.

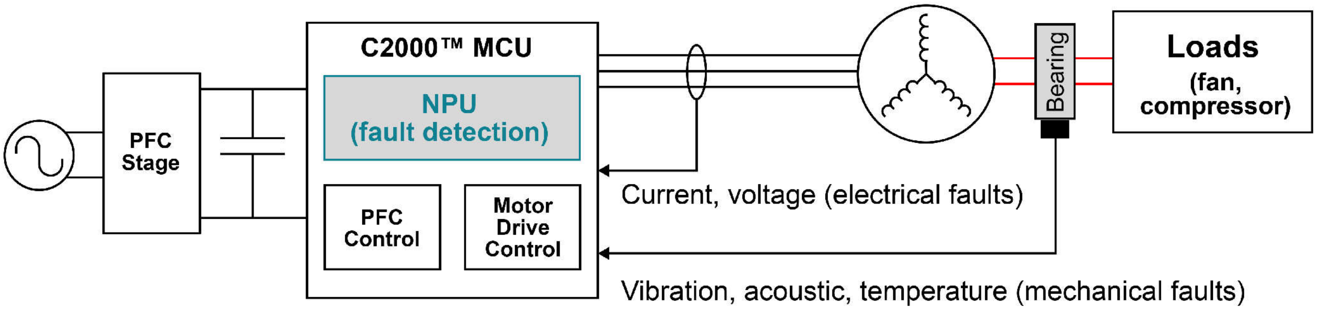

Figure 3. *Edge AI-enabled fault monitoring solution in a real-time control system.*

NPUs are a type of dedicated hardware accelerator that can be integrated into the design of an embedded processor, enabling the level of computation required. For example, TI's TMS320F28P55x series real-time C2000 MCUs like the **TMS320F28P550SJ** feature an on-chip NPU for motor bearing and solar arc fault detection, offloading these processes from the main CPU, which is used for real-time motor control. Detecting faults early helps reduce downtime and repair costs, while also improving safety. Motor fault detection is crucial in applications such as heating, ventilation ,and air conditioning (HVAC); factory automation; robotics; and electric vehicles. These applications can use time-series data aggregated from various sensors to make accurate inferences and rapidly detect anomalies in a system. Read the technical article, "**Optimizing system fault detection in real-time control systems with edge AI-enabled MCUs**," to learn more about how edge AI-enabled C2000 MCUs simplify motor and arc fault detection.

For perception-based applications like those for vision processing, microprocessors need to include the necessary components of a vision pipeline, along with NPUs capable of meeting the AI performance required. TI's analytics processors like the **AM67A** and the **TDA4VM** feature integrated deep learning accelerators to optimize for AI performance as well as power consumption. These deep learning accelerators are each composed of a C7x DSP with an attached matrix-multiply accelerator that significantly improves the processing performance on neural networks common in vision processing. The architecture leverages TI's DSP technology to achieve efficient data movement and optimize power consumption by centralizing AI computation onto homogenous computing elements.

To ensure embedded processors are as easy to design with as possible, semiconductor manufacturers like TI also work closely with **third-party partners** that develop "ready-to-deploy" hardware components like a system-on-module. These hardware solutions are designed to support specific embedded processors and core components on a single PCB.

Edge AI software innovations: Along with the advancements in efficient AI computation in embedded devices from a hardware perspective, 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 to enhance one's AI expertise.

For designers who have more familiarity with AI models, open-source platforms such as PyTorch and TensorFlow can help simplify the process of developing edge AI solutions by abstracting away many of the specifics of each embedded platform. It's possible to import models developed using these tools to compatible embedded devices.

For developers who want to further streamline their testing and deployment of AI models on TI devices, **TI's Edge AI Studio**, a collection of web-based tools, is available. Edge AI Studio was designed to simplify and accelerate the development of edge AI applications on TI embedded devices using remote TI hardware and a graphical user interface. 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, and 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 vision processors 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, it's 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 at **https://www.ti.com/technologies/edge-ai.html**.
- Learn about TI's edge AI vision processors in the white paper, **Designing an Efficient Edge AI System With Highly Integrated Processors**.
- Read about key design challenges for smart home and retail edge AI vision-based applications in the technical article, **How Vision Processors Are Expanding Edge AI Capabilities in Video Doorbell and Smart Retail Designs**.

IMPORTANT NOTICE AND DISCLAIMER

TI PROVIDES TECHNICAL AND RELIABILITY DATA (INCLUDING DATA SHEETS), 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 will 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 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.

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

Mailing Address: Texas Instruments, Post Office Box 655303, Dallas, Texas 75265