Embedded Artificial Intelligence: Intelligence on Devices

Hsiao-Ying Lin, Huawei France

Embedded artificial intelligence (AI) seamlessly integrates AI into everyday devices. By bringing AI closer to the data source, embedded AI empowers real-time decision making, enhanced efficiency, and new possibilities across diverse domains.

DISCLAIMER

This article contains the views of the author. The opinions expressed here are hers alone.

Digital Object Identifier 10.1109/MC.2023.3280397
Date of current version: 23 August 2023

Imagine a typical day in modern society. People wake up in the morning and check how well they slept at night using their smartwatches. They then ask their voice assistants about the weather to choose their outfits. They activate smartphones with facial identity scanners and look up their daily schedules. Finally, they drive to their offices using vehicles with autonomous driving functions. Embedded artificial intelligence (AI) will be widespread.

EMBEDDED AI

Embedded AI is the integration of AI into resource-limited devices or systems, such as wearable devices, smartphones, smart home devices, industrial automation systems, robotics, and autonomous vehicles, just to name a few. This is also called *on-device AI* or *TinyML*. According to the TinyML Foundation[1]:

> "Tiny machine learning is broadly defined as a fast-growing field of machine learning technologies and applications including hardware, algorithms, and software capable of performing on-device sensor data analytics at extremely low power, typically in the [milliwatt] range and below, and hence enabling a variety

0018-9162/23©2023IEEE

EDITOR **HSIAO-YING LIN**
IEEE Member; hsiaoying.lin@gmail.com

of always-on use-cases and targeting battery operated devices."

AI algorithms enable these devices to perform tasks efficiently, accurately, and autonomously. Unlike cloud-based AI services, which rely on cloud-based computing and data transfer, embedded AI implements real-time data analysis of devices and delivers reactions via the devices. According to the Maximize Market research,[2] the global embedded AI platform market is expected to grow by 5.4% per year, reaching a market value of US$45.51 billion by 2029, where the expected consumers include the healthcare and automotive sectors.

Similar concepts, such as edge AI, AI of Things (AIoT), and embodied AI, exist. Edge AI is the integration of AI on devices at the network edge where these devices have networking capabilities. AIoT is the integration of AI with Internet of Things ecosystems and devices, which have a broader range than a single device. Embodied AI is the integration of AI and technologies with robotic or physical systems so that these systems can perceive, interact with, and navigate the physical world. Although embedded AI, edge AI, AIoT, and embodied AI overlap, they have different focuses.

APPLICATIONS OF EMBEDDED AI

Embedded systems are specialized computing systems designed for specific tasks and applications. They feature low-power processors with clock speeds ranging from a few megahertz to several gigahertz, tailored to meet the requirements of the targeted application while optimizing the power efficiency. Figure 1 shows some example domains of embedded AI that present the flavor of embedded AI.

- *Smartwatch/fitness tracker:* Real-time activity recognition is an example of embedded AI in a smartwatch or fitness tracker. The embedded AI algorithms in these devices can analyze sensor data, such as accelerometer and heart rate measurements, to accurately identify various physical activities (i.e., walking, running, cycling, or swimming). This enables the device to provide users with personalized activity tracking, feedback, and insights.
- *Medical devices:* The U.S. Food and Drug Administration (FDA) has created and maintained a list of cleared/approved AI/machine learning (ML)-enabled medical devices.[3] One example is the AI-powered whole-breast ultrasound system cleared by the FDA in 2022.[4] The system contains a portable and wearable ultrasound scanner. The scanner captures the entire breast volume without the need for a trained ultrasound operator and offers 3D visualization of the breast tissue. The system provides physicians with a set of AI/ML-enabled tools for decision making and patient management.
- *Autonomous drones:* An example of embedded AI in autonomous drones is obstacle avoidance.[5] Embedded AI algorithms enable autonomous drones to analyze sensor data, such as camera feeds, in real time to detect obstacles, people, or other aircraft in their flight path and decide the flight trajectory to avoid collisions and navigate safely in complex environments.
- *Smart speakers:* Keyword spotting in smart speakers involves the use of embedded AI algorithms to recognize specific keywords or wake words from audio inputs. The embedded AI system allows efficient and real-time keyword detection and enhances the overall user experience by enabling hands-free interaction with a smart speaker.
- *Smartphones:* Facial recognition is a typical embedded AI application embedded in smartphones. Embedded AI algorithms utilize the front-facing camera of the device and advanced computer vision techniques to analyze and identify the unique facial features of the user. They can be implemented in smartphone-unlocking and payment applications.
- *Autonomous driving:* Object detection is an example of embedded AI in autonomous driving. Using various sensors (i.e., cameras, lidar, and radar), the embedded AI system analyzes the data in real time and identifies and classifies objects, such as vehicles, pedestrians, traffic signs, and road markings. This information is crucial for making decisions and controlling vehicle movements, enabling autonomous driving functions, such as adaptive cruise control, lane-keeping assistance, automatic emergency braking, and object avoidance.

MERITS AND METRICS OF EMBEDDED AI

As cloud-based AI applications, such as ChatGPT, are emerging, embedded AI

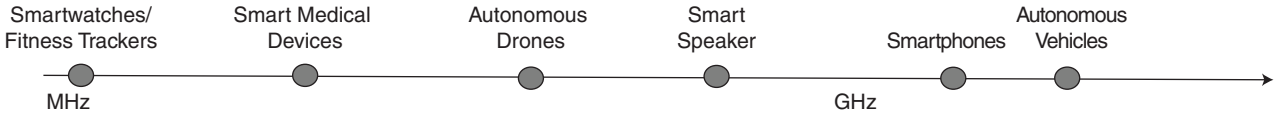

FIGURE 1. Embedded AI example domains according to the approximated clock speeds.

offers several benefits over cloud-based AI in various domains.

- *Bandwidth efficiency:* Embedded AI reduces the reliance on cloud connectivity, resulting in reduced data transmission and lower bandwidth requirements. This is beneficial in scenarios with limited or expensive network connectivity because it optimizes the bandwidth usage and reduces costs.
- *Energy efficiency:* By performing local computations on devices, embedded AI minimizes the need for data transmission to and from the cloud, resulting in reduced energy consumption. This is particularly important for battery-powered devices.
- *Reduced latency:* Embedded AI minimizes the latency associated with data transmission over networks by processing data locally on devices. This is important for time-sensitive applications, such as the collision avoidance of autonomous vehicles.
- *Privacy:* Embedded AI allows data to be processed locally without transmitting sensitive information to external servers. This enhances data privacy as the data remain within the device, thereby reducing the risk of data breaches.

To evaluate the performance of embedded AI, MLperf Tiny[6] is an available benchmark suite tool specifically designed to evaluate the performance and efficiency of ML inference on resource-constrained devices. It offers standardized metrics and testing methodologies that cover a range of tasks, including image classification, object detection, and keyword spotting; additionally, it enables comparisons between different platforms and implementations of embedded AI applications. The metrics include:

- *Inference latency:* This measures the time it takes for the ML model to process a single input and produce an inference result. It is measured in milliseconds.
- *FPS:* The FPS measures the number of inferences that the ML model can perform per second.
- *Accuracy:* This measures the correctness of the predictions of the ML model compared with the ground truth or expected outputs. This represents the performance of the model for a given task.
- *Power efficiency:* This measures how effectively the ML model utilizes the device's power resources, often in terms of the number of inferences per watt (inferences/watt).

TECHNICAL ENABLERS

Hardware accelerators, software toolchains, and deep neural network optimization technologies are the three main pillars to enable embedded AI, and their relationships are illustrated in Figure 2. The interplay between them contributes to the success of the implementation of embedded AI.

AI hardware accelerators are designed at the hardware level to accelerate the AI training and inference processes. To facilitate AI hardware accelerators, dedicated software toolchains are required for training and compiling tasks, where the training process seeks an AI model within the resource budget, and the compiling process converts the model into a target device with customized instructions and memory constraints. A summary of the selected available AI hardware accelerators and their software development tools for reference is shown in Table 1.

In addition to specialized AI hardware accelerators and their software tools, deep neural network optimization techniques are critical enablers of the success of embedded AI. However, this topic remains an active area of research. Model compression is one of these optimization techniques and there are several techniques for model compression.

- Network pruning includes neuron and edge pruning. Neuron pruning disables certain neurons by removing them from the model. In principle, edge pruning removes edges. However, in practice, edge pruning is implemented for edges with weights of zero.
- Knowledge distillation uses a (larger) teacher model to transfer knowledge to another student model of smaller size.
- Parameter quantization is another type of model compression technique. While the major numerical format for model weights is a 32-bit float, model weights or activations are represented in lower bit widths, such as 16-bit, 8-bit, 4-bit, 2-bit, and 1-bit. This is achieved by truncating the weights or activations of the trained models. This can also be conducted during the training phase when the training algorithms are designed in a sophisticated manner.

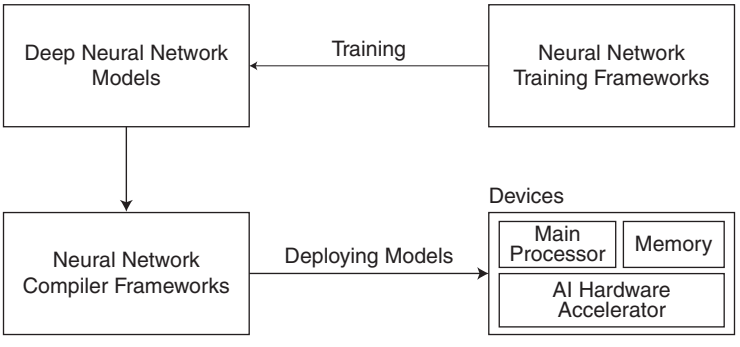

FIGURE 2. Abstract workflow of embedded AI.

› Dynamic computation aims to train the models in multiple environments. The available techniques include dynamic depth (where the final layers are designed to increase the model accuracy), slimmable neural networks[7] (dynamic width where parallel neurons are increased for model accuracy), SkipNet,[8] and runtime pruning.

Embedded AI brings real-time AI capabilities directly to devices and systems. It offers efficiency, reduced latency, and privacy compared with cloud-based AI systems. Because of its potential to revolutionize industries and enable intelligent and autonomous systems, embedded AI paves the way for innovative applications and empowering devices to make intelligent decisions closer to where they are required.

TABLE 1. Examples of hardware and software for embedded AI.

AI hardware accelerators	Software development tools
Google coral edge tensor processing units	TensorFlow Lite is a lightweight version of the TensorFlow framework specifically designed for mobile and embedded devices.
Apple Bionic chips	Apple Core ML is a framework that allows developers to integrate ML models directly into iOS, iPadOS, macOS, and watchOS devices.
ARM ML processor	Arm cortex microcontroller software interface standard is a software library that offers a collection of efficient and standardized functions for developing embedded applications on Arm Cortex-M microcontrollers
Intel Movidius visual processing unit	Intel open visual inference and neural network optimization (OpenVINO) allows for optimizing and deploying pre-trained deep learning models across Intel architectures.
NVIDIA Jetson	NVIDIA JetPack includes libraries, tools, and frameworks for developing AI applications on NVIDIA's Jetson platform.
Meta training and inference accelerator (MTIA) application specified integrated circuits	PyTorch Mobile is a lightweight version of the PyTorch framework that enables deploying AI models on mobile and embedded devices.
Imagination neural network accelerators (NNAs)	Imagination Neural Compute-SDK supports heterogeneous compilation across CPU, GPU, and NNA and various quantization tools.
Mediatek AI processing unit	MediaTek Neuropilot ML kits provide converter, quantization, and network compression tools.
Baidu Kun Lun Core	Kun Lun Core software development kit includes Kunlun Core's driver, virtualization module, and its runtime library.
Kneron edge neural processing unit	Kneron software development kits include tools to build firmware for Kneron AI system-on-chip family with application examples.

REFERENCES

1. tinyML Foundation. Accessed: May 19, 2023. [Online]. Available: https://www.tinyml.org/
2. "Embedded AI computing platforms market: Global industry analysis and forecast (2022-2029)," Maximize Market Research, Narhe, Pune, India, Feb. 2023. Accessed: May 17, 2023. [Online]. Available: https://www.maximizemarketresearch.com/market-report/global-embedded-ai-computing-platforms-market/111905/#details
3. "Artificial intelligent and machine learning (AI/ML)-enabled medical devices," U.S. Food and Drug Administration, Silver Spring, MD, USA, Oct. 2022. Accessed: May 19, 2023. [Online]. Available: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices
4. A. Park. "Wearable, AI-powered whole-breast ultrasound system cleared by the FDA." Fierce Biotech. Accessed: May 19, 2023. [Online]. Available: https://www.fiercebiotech.com/medtech/wearable-ai-powered-whole-breast-ultrasound-system-cleared-fda
5. H. Kesteloo. "Skydio's obstacle avoidance and self-flying capability explained." DroneDJ. Accessed: May 19, 2023. [Online]. Available: https://dronedj.com/2019/10/01/skydios-obstacle-avoidance-self-flying-capability/
6. C. Banbury et al., "MLPerf Tiny benchmark," in *Proc. Neural Inf. Process. Syst. Track Datasets Benchmarks*, 2021, pp. 1–15.
7. J. Yu, L. Yang, N. Xu, J. Yang, and T. Huang, "Slimmable neural networks," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2019, pp. 1–12.
8. X. Wang, F. Yu, Z.-Y. Dou, T. Darrell, and J. E. Gonzalez, "SkipNet: Learning dynamic routing in convolutional networks," in *Proc. Eur. Conf. Comput. Vision (ECCV)*, 2018, pp. 409–424.

HSIAO-YING LIN is a principal researcher at Huawei France, 92100 Boulogne-Billancourt, France. Contact her at hsiaoying.lin@gmail.com.