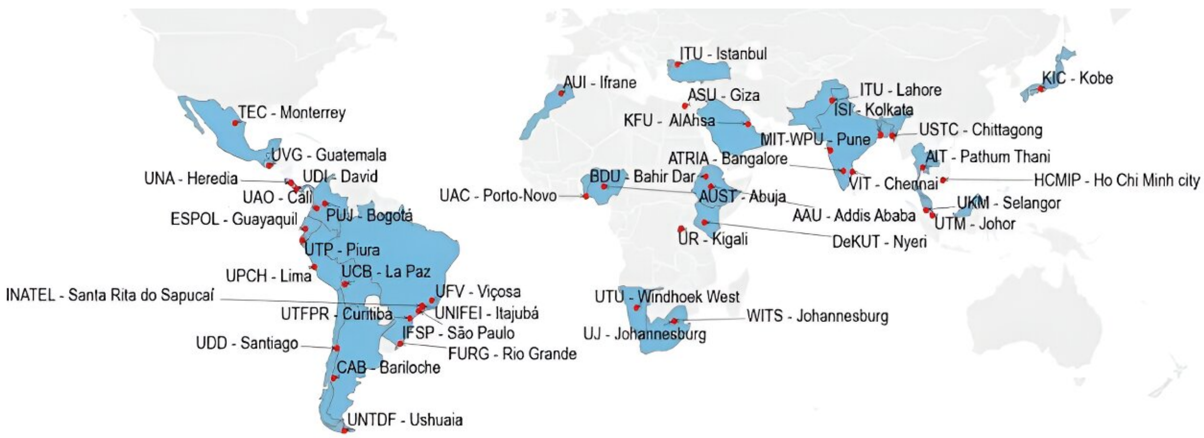
AI in the developing world: How 'tiny machine learning' can have a big impact

February 10 2024, by Marco Zennaro, et al.



Map of the TinyML Academic Network. More than 50 universities are part of the network as of February 2024. Credit: Marcelo Rovai, CC BY-SA

The landscape of artificial intelligence (AI) applications has traditionally been dominated by the use of resource-intensive servers centralized in industrialized nations. However, recent years have witnessed the emergence of small, energy-efficient devices for AI applications, a concept known as tiny machine learning (TinyML).

We're most familiar with consumer-facing applications such as Siri, Alexa, and Google Assistant, but the limited cost and small size of such devices allow them to be deployed in the field. For example, the technology has been used to detect mosquito wingbeats and so help prevent the spread of malaria. It's also been part of the development of low-power animal collars to support conservation efforts.

Small size, big impact

Distinguished by their small size and low cost, TinyML devices operate within constraints reminiscent of the dawn of the personal-computer era—memory is measured in kilobytes and hardware can be had for as little as US$1. This is possible because TinyML doesn't require a laptop computer or even a mobile phone. Instead, it can instead run on simple microcontrollers that power standard electronic components worldwide. In fact, given that there are already 250 billion microcontrollers deployed globally, devices that support TinyML are already available at scale.

A number of development packages for TinyML applications are available. Two popular options are Arduino and Seeed Studio, both of which come with additional sensors for audio, vision, and motion-based applications.

How does it work?

Like classical machine learning, TinyML involves data collection—often from Internet of Things (IoT) devices—and cloud-based training. Let's consider an outdoor object-detection application—for example, counting the number of cars on a street to see how heavy the traffic there is. In the classical ML process, images have to be gathered using a webcam and sent to a cloud server where the training takes place. Once the trained model provides an acceptable level of accuracy, the system is ready to detect cars from a new video feed. The ML model runs on the cloud, so an Internet connection is necessary.

In the TinyML system, however, the model is deployed on the device itself and is ready to detect objects with no need for connectivity. The first part of the process (gathering data and training the model on the cloud) follows the classical ML model but the inference phase (detecting objects) runs on the device itself. This is how TinyML diverges from traditional server-based architectures: it deploys pre-trained compact models optimized for limited resources onto embedded devices, enabling real-time, low-power data analysis and decision-making, all independent of cloud connectivity.

TinyML offers several advantages over traditional centralized server-based models:

- Affordability: the technology's low cost makes these devices accessible to a wide range of users including educational institutions and students in the developing world.
- Sustainability: the modest energy consumption produces a low carbon footprint, reducing impact on the environment.
- Flexibility and scalability: it enables the development of applications that address the needs of local communities rather than global agendas.
- Internet independent: Because everything is embedded, TinyML devices can operate without online connectivity. This is particularly beneficial for the third of the world that still does not have Internet access.

TinyML applications already power personalized sensors for athletics and provide localization where GPS isn't available. They're also employed by startups such as Useful Sensors, which offers privacy-conserving conversational agents, QR code scanners, and person-detection hardware. Only through the use of TinyML could these smart devices run on the low-cost, low-power microcontrollers.

Developing in the Global South

To help the use of TinyML grow in regions where a centralized machine-learning model would face significant challenges, we built TinyML4D, a network of academic institutions in developing countries. It already includes more than 40 countries spanning the Global South from Columbia to Ethiopia to Malaysia.

With support from UNESCO's International Centre for Theoretical Physics (ICTP) and from Harvard University's John A. Paulson School of Engineering and Applied Science, the network was launched in 2021. Its aim is to develop a community of educators, researchers and practitioners focused on both improving access to TinyML education, and developing innovative solutions to address the unique challenges faced by developing countries.

To make all this possible, we needed to develop ways to share educational resources globally. Initial efforts included distributing TinyML hardware kits to selected universities with budgetary challenges. We also organized global and regional (Africa, Latin America, and Asia) workshops and training sessions. Using a mixture of in-person, online and hybrid methods, we've reached more 1,000 participants in over than 50 countries. The combination of no-cost or low-cost hardware resources, combined with open-source course materials and workshops has enabled TinyML to be taught by many of our network members in their home countries.

Beyond our workshops and training activities, we have launched a series of regional collaborations, outreach activities and virtual "show and tell" events to share best practices and augment our network's impact among practitioners. Throughout, there has been a strong focus on addressing the United Nations' sustainable development goals (SDGs).

These collaborations have led to multiple peer-reviewed papers on TinyML applications. In addition to the solution to detect mosquito species, which could lead to more efficient malaria-control campaigns, others include the responsible use of intelligent sensors and low-cost solutions to monitoring atrial fibrillation and sinus rhythm. They're also used by Cornell University's "Elephant Listening Project" as well monitoring water quality in aquaculture to help make it more sustainable, a project supported by EU's Horizon 2020 program.

Looking forward

TinyML represents a transformative approach to artificial intelligence and is especially pertinent to developing countries. It offers a sustainable path toward democratizing AI technology, fostering local innovation, and addressing regional challenges.

The growth of TinyML devices and applications is not without potential challenges and risks, however. The number of applications and devices is expected to rise from the millions shipped today to 2.5 billion devices in 2030, and that could lead to increased electronic waste due to the low-cost nature of devices. There's also the risk of embedded biases in critical ML models—because they operate standalone, there's no option for updates. Finally, there are privacy concerns due to the discrete integration of devices in the environment. As the field evolves, it will be crucial to navigate these issues responsibly, and so help ensure that TinyML remains a tool for positive change and sustainable development.

Provided by The Conversation