Electronic Design

ALAN R. EARLS, Technology Writer

Can tinyML Bring Machine Learning to the Masses?

Delivering smarter smarts to edge and free-standing devices, tinyML can be part of a broader AI/ML deployment.

In the ebb and flow of ideas and innovation, definitions and assumptions are always evolving, e.g., mini vs. micro. So, now the expanding realm of AI and machine learning has a new star, the world of "tiny" machine learning. It consists of very low-power hardware and software that provides on-device analytics of sensor data.

This stands in sharp contrast to much artificial-intelligence/machine-learning (AI/ML) activity, which typically runs on power-hungry processors and is performed away from the realm of the sensors.

Central to many ML systems is a model, typically "trained" on representative data that will make a good basis for then digesting other data, refining what it "knows" and enabling it to accomplish useful tasks. Machine-learning systems of this type are present all around us, for instance in social-media functions, search engines, and spam filtering.

Most of these common ML applications require significant computing resources, often cloud-based. This requirement for processor cycles in general-purpose computing resources has meant that ML applications haven't been deployed as widely as they could have been, driving a quest for ML to able to run on lower-resourced devices and systems. These efforts are usually lumped under the term "Tiny Machine Learning" or "tinyML."

A tinyML Summit in March 2019 helped launch a community, eventually involving some 90 companies. The timing was related to the fact that hardware was becoming sufficient to support ML on a much smaller scale. The need for such systems was already clearly established.

Progress was made, and continues, on algorithms, networks, and models as small as 100 kB or less. This meant that real-world applications could be conceived in vision and audio areas. It also meant that tinyML could be applied in edge applications, enabling better responsiveness and, incidentally, greater intelligence at this level.

The open-source Frog Sensor (top) from Ribbit Network is built around a Raspberry Pi CM4 module. (Ribbit Network)

By deploying at the edge, the challenge of latency is reduced or eliminated. Without the uncertainty regarding responsiveness imposed by distance, many more tasks can be exposed to ML techniques. Being at the edge delivers a more predictable approximation of real-time behavior.

Emerging tinyML Applications

Applications for tinyML are starting to appear in the real world, including in farming and sustainability. One, developed by Niolabs, which works on Internet of Things (IoT) challenges, aimed to sharply improve the way water is used in farming. Although sensor technologies already exist to assess soil moisture, sunlight, and other factors, centralizing information and making appropriate decisions for all crops in each part of a farm was proving difficult.

tinyML provided a solution by allowing microprocessors, with access to hyperlocal information, to make optimal water-use decisions, thereby ensuring successful crops with minimal water use.

Another tinyML application from Ribbit Networks yielded the Raspberry Pi CM4-based Frog sensor *(see figure)*, which monitors local carbon-dioxide levels to complement data provided by satellites. The open-source design even has a tutorial showing how to build your own tinyML system.

Many existing tools and techniques can be applied to create tinyML applications, but tinyML frameworks can help. Some of the best-known frameworks for getting more out of microprocessors are PyTorch Mobile, Edge Impulse, and TensorFlow Lite (TF Lite):

- ***PyTorch Mobile*** is a runtime beta release that helps transition from training a model to deployment, while staying within the PyTorch ecosystem. According to the organization, it provides an end-to-end workflow that simplifies the research-to-production environment for mobile devices. In addition, it paves the way for privacy-preserving features via federated learning techniques.
- ***Edge Impulse***, an embedded machine-learning company, offers the ability to "build datasets, train models, and optimize libraries to run on any edge device, from extremely low-power MCUs to efficient Linux CPU tar-gets and GPUs."
- ***TensorFlow Lite for microcontrollers*** (TF Lite), which grew out of the Google-originated TensorFlow, is also widely cited for use in edge devices. It can deliver functionality using as little as a few hundred kilobytes of data.

What else can be done with tinyML technology? A lot. One area with huge promise is predictive maintenance (the focus of a recent tinyML foundation event), where conditions such as vibration and temperature or even general sound monitoring are already being monitored to feed ML systems—think wind turbines. Those systems can be smart enough to schedule maintenance or report degrading conditions to support staff for further diagnostics or repair.

In a worst-case situation, a tinyML system could shut down a machine to avoid further damage. A tinyML system that monitors motor sounds, developed using Edge Impulse, is described here. This system integrates a microphone and processor in one device.

tinyML has also contributed to applications in areas such as customer experience, delivering personalization at the edge, and predictive maintenances for industrial settings.

Other Possible Benefits Brought by tinyML

Some overarching plusses are potentially available through deploying tinyML. In addition to reducing latency, tinyML applications typically require much less power than traditional approaches. The devices themselves typically use very little power and because there's less communication with, or processing by, a traditional CPU, it substantially reduces the demand for power. A single instance will hardly make a difference, but broad tinyML deployment can potentially yield significant savings.

While data privacy and security concerns may arise with the device itself, some of those can be addressed by minimizing data storage. And the centralized storage of data used traditionally is also sharply reduced.

All in all, the "tiny" in tinyML offers lots of potential.

References

"What is Machine Learning?," IBM.
"Practical Applications of tinyML," Pete Warden, YouTube.