TinyML and AI on the Edge: Can Machine Learning Fit into 256 Kilobytes?

Uli Hitzel
Executive Geek

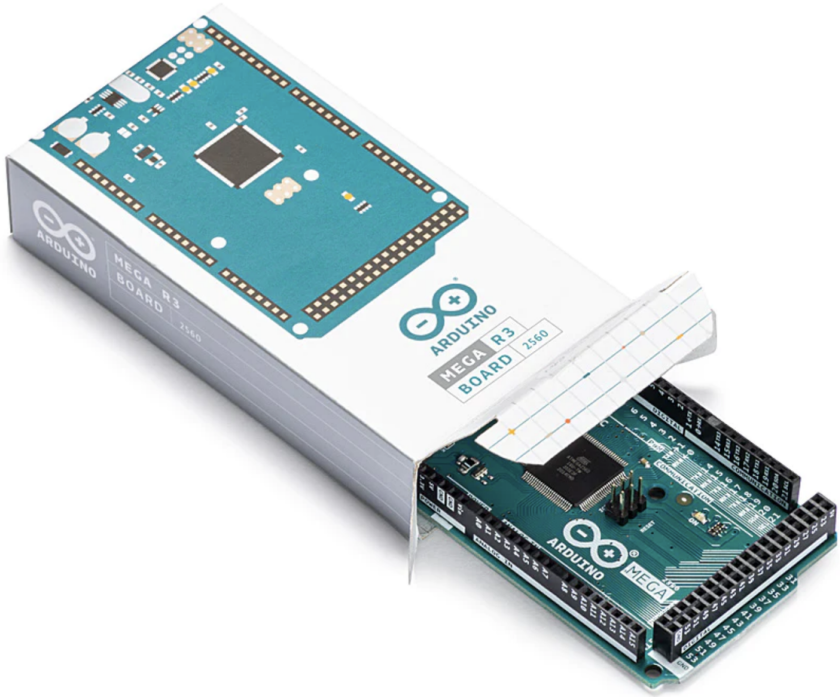
June 29, 2023

We hear a lot about training and deploying machine learning models on massive data in the cloud, but what about putting ML to work in resource-constrained environments - such as the family fridge or a self-driving car where every millisecond counts? Let's explore the emerging field of TinyML and what it means for AI at the edge.

When we talk about Artificial Intelligence and Machine Learning, we immediately think of huge datasets, and massive compute infrastructure running on powerful machines in the cloud. According to a **Forbes article**, advanced language models like OpenAI's Chat GPT cost millions of dollars to operate because of the large amounts of compute power and memory needed. This makes sense, considering that the language model GPT-4 has more than a trillion parameters. We're talking really big numbers here, if that one really is true.

However, while cloud-based AI has tremendous applications and potentials, it is not the only way. We have seen more and more AI applications reaching our end users, from Alexa and Siri to autonomous vehicles. Even areas such as medical care and manufacturing are benefitting from AI on the edge, where decisions need to be made quickly and cost effectively. But how is it possible to fit such powerful machine learning models into low-power and resource-constrained environments, such as chip-sized IoT (Internet of Things) devices or embedded systems with capacity for just 256 kilobytes (kB) of RAM?

Deploying machine learning models in these ultra-low power environments can be extremely challenging due to the limited compute resources, memory, and power availability. Not to mention the real-time constraints of some AI-enabled use cases where latency must be minimized as much as possible.



The answer to this challenge is a new and emerging field called TinyML, which is designed to bring machine learning models to the edge. TinyML is a special, highly optimized form of Machine Learning that can run on very small and low-powered computing devices, such as embedded systems, or even those powered by batteries. It consists of using the right model architecture, algorithms, and techniques for the given use case, such as weight pruning and quantization, to fit the model into the constrained environment.

The key to TinyML is to take the model which would normally be huge in size, and distilled it into the appropriate size for the device using it. Data scientists and engineers use a variety of techniques including weight pruning, weight quantization, and model compression to reduce the size of a model, while still maintaining accuracy and performance.

In distributed training, engineers often use large machine learning models that use multiple GPUs, CPUs, or TPUs (Tensor Processing Units) in the cloud. This helps to generate large datasets which are usually difficult or expensive to produce in physical experiments. On the other hand, model distillation takes the large model and distils it to fit into the specific platform (TinyML application) using techniques such as weight pruning, weight quantization, and model compression. Model distillation enables a much smaller model to produce the same results, without sacrificing accuracy, while using a fraction of the compute capacity.

Let's take a look at some of the use cases in which TinyML is already being put to work. Industrial predictive maintenance, for example, uses tiny low-powered systems to continuously monitor machines for potential faults, leading to cost saving for companies. The way this works is by attaching a magnetic sensor to the equipment, and then running ML

models on the device. This way, potential issues could be spotted before they even happen, and no data has to be sent to the cloud in order for this to happen.

In an electric vehicle, for example, TinyML can be used to improve the autonomous driving systems. Tiny low-power sensors and cameras are attached to the vehicle, and the ML models run in the car itself. This way, critical decisions can be made in real-time without having to wait for streaming data over the network, and AI models can still run with minimal latency despite the low power capacity.

In the healthcare space, we see TinyML being used in autonomous mosquito traps that detect mosquitoes and prevent the spread of diseases such as dengue and malaria.

In agriculture, an app called Nuru helps farmers to detect diseases in plants by simply taking a picture without having to rely on an internet connection. You can see from these examples of how TinyML can fit into almost any environment, and how it is enabling us to put powerful machine learning models to work in cases where it was not previously possible.

The TinyML space is constantly growing with new and exciting use cases. It is predicted that more companies will start to adopt this, as their devices become connected to the internet. Companies such as Microsoft, Google, Amazon, among others, are already developing unique ML solutions for small, low-powered machines, such as the AWS DeepLens and Google Coral. We are also seeing the emergence of new startups working on TinyML solutions.

We can assume that in the near future, ML models will become more complex and powerful, while also becoming more lightweight and efficient. We will also see more edge deployments, with devices such as sensors, controllers, and actuators, and even smaller and more affordable devices such as wearables, contact lenses, and even sub-dermal implants.

AI on the edge is the key to unlocking the potential of machine learning technology in resource-constrained devices, and it is bringing us closer to a world where AI can be applied to solve a variety of problems, from healthcare and manufacturing to robotics and autonomous vehicles. Check out the course "**Fundamentals of TinyML**" on edX!