Why the Future of Machine Learning is Tiny and Bright

A New Paradigm

Over the past decade, we have witnessed the rise of machine learning and its gradual incorporation into the industry. More recently, we have begun to witness the bifurcation of machine learning into two separate paradigms based on scale. The first of these is **cloud-based machine learning**, working at a large scale and utilizing vast swathes of data stored in data lakes alongside scalable frameworks such as Dask or PySpark. Cloud-based machine learning has been the major focus of industry and is largely compute-centric; centralizing data in a single location before performing computing operations.

In contrast, **microcontroller-based machine learning**, often referred to as TinyML, instead focuses on the small scale and is inherently data-centric. Embedding machine learning methods within microcontrollers provides new opportunities that have the potential to offer more efficient computation, reduced communication overhead, cost savings, as well as localized computation.

Here, we will outline some of the key points that make TinyML an important player in the future machine learning ecosystem.

Machine Learning at the Edge

Computing data locally, rather than streaming data to the cloud or an edge server, is a cheaper and more power-efficient solution for intelligently processing data available on endpoint devices from attached peripherals (e.g., sensors). This not only conserves energy by limiting communication overhead but also unlocks new applications and use cases. Consequently, there is a growing interest in providing machine learning functionality for endpoint devices. In recent years, it has become possible to take noisy signals like images or audio data and extract meaning from them using neural networks, with TinyML allowing us to run these networks on our microcontroller devices. Since sensors themselves use little power, TinyML can be used to perform real-time machine learning computation on our devices without the need for external communication. We can see applications of this in our daily lives; both Apple and Google run always-on machine learning (ML) based neural networks for voice recognition on these kinds of power-efficient chips (think “OK Goole” or “Hey Siri”).

Tiny Computers Are Ubiquitous

The shift towards microcontroller-based machine learning is largely a result of the increasing number of internet-connected endpoint devices. Whereas ten years ago people had at most a phone and laptop, nowadays, a single person can have a dozen or even more personal devices, ranging from smart thermostats and TVs to smart watches and e-readers. Figure 1 (data from IC Insights) suggests that there will be over 30 billion microcontrollers sold this year, with that number increasing year by year -- to put that into perspective, there are only about 10 million servers in use across the planet and only 4 billion smartphone users. Microcontrollers typically do not get much attention because they are often used to replace functionality that older electro-mechanical systems could do, in cars, washing machines, or remote controls, but they are ubiquitous.

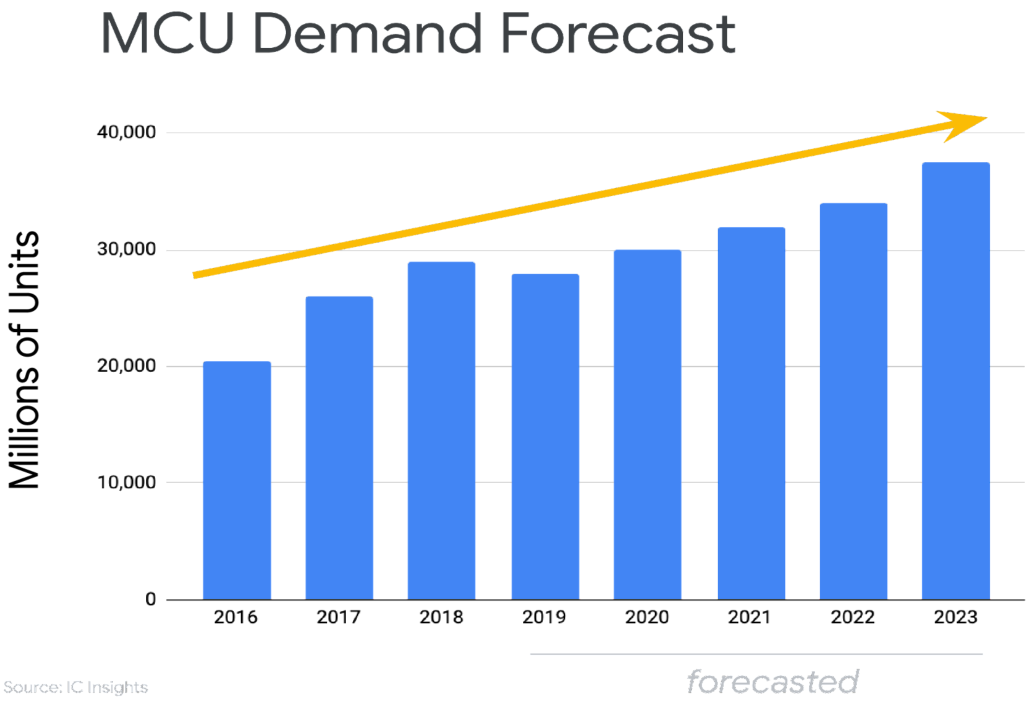
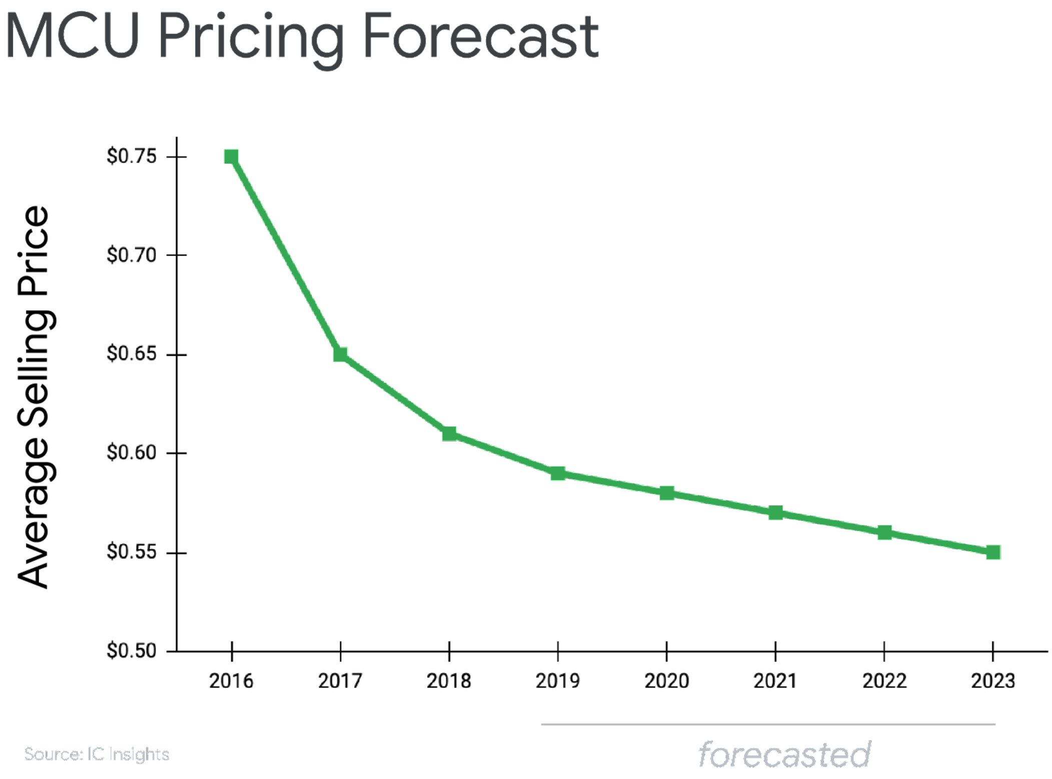


FIGURE 1

Microcontrollers are Cheap

Microcontrollers are much cheaper than conventional hardware such as phones and laptops, because they aren't designed to be general-purpose computational workhorses that run complex workloads. Instead, they are often designed to perform specific tasks, slowly, albeit steadily. The average price of a microcontroller unit (MCU) is already less than USD $1. Figure 2 shows the average sales price (ASP) for a microcontroller is hovering close to 55 cents. As demand continues to rise, it is expected that this will soon fall below 50 cents. On a global scale, the microcontroller market ($M) is poised to grow by USD $6.74 Billion between 2020-2024.



Source: IC Insights

FIGURE 2

MCUs are Resource-Constrained, Ultra-low Power Systems

The holy grail for almost any embedded device is for it to be deployable anywhere and require no maintenance like docking or battery replacement. Any device that requires tethered electricity faces a lot of deployment barriers. It can be restricted to only places with electrical wiring. Even where electrical wiring capability is available, it may be challenging for practical reasons to plug something new in, for example, on a factory floor or in an operating theatre. Putting something high up in the corner of a room means running a cord or figuring out alternatives like power-over-ethernet. The electronics required to step down or convert the main-line high voltage to low voltage are expensive and waste energy.

But perhaps the most significant barrier to achieving ubiquitous deployment of computers everywhere is how much energy an electronic system uses. Here are some rough numbers for standard components based on figures from Smartphone Energy Consumption, and so it is no wonder that smartphones need to be tethered to the wall and recharged every night:

- Active cell radio might use 800 milliwatts.
- A display might use 400 milliwatts.

- GPS is 176 milliwatts.
- Gyroscope is 130 milliwatts.
- Bluetooth might use 100 milliwatts.
- Accelerometer is 21 milliwatts.

A key opportunity with using MCUs is they require a very minimal amount of energy. A microcontroller itself **might only use a milliwatt or even less**. Still, you can see that peripherals (accelerometer, gyroscope, GPU, etc.) require much more energy to stay on.

A coin battery might have 2,500 Joules of energy to offer, so even something drawing only one milliwatt will have severe consequences as that means it can run continuously for only about 30 days. Of course, most current products use “duty cycling” and power naps to avoid being always on, but we see how tight the budget is even then. The overall thing to take away from these figures is that while processors and sensors can scale their power usage down to microwatt ranges, displays and especially radios are constrained to much higher consumption, with even low-power wifi and Bluetooth using tens of milliwatts when active. The physics of moving data around is generally well-known to require a lot of energy.

The general wisdom is that the energy an operation takes is proportional to how far you have to send the bits. CPUs and sensors send bits a few millimeters and are cheap. Radio sends them meters or more and is expensive. We don’t see this relationship fundamentally changing, even as technology improves overall. We expect the relative gap between computing and radio costs to get more expansive because we see more opportunities to reduce computing power usage.

Tiny Machine Learning (TinyML) on MCUs

Can you imagine a future where every one of the 250B tiny MCUs runs intelligent machine learning algorithms that can sense their surroundings, predict events in real-time, and even make nudges or recommendations based on the sensors’ activity? We can transform the world.

For instance, smart consumer devices like smart toothbrushes will have MCUs that are tightly coupled with sensors that dynamically adjust to your brushing intensity based on pressure. Medical devices in the future may use microcontrollers with biomedical sensors to control drug delivery as and when needed. Microcontrollers in automotive applications can aid functional safety on the road, detecting engine conditions, etc. to ensure our safety. Finally, MCUs will likely have widespread applications in industrial devices for continuous process monitoring and anomaly detection for applications such as predictive maintenance that can save millions of dollars in productivity loss and downtime by forewarning maintenance engineers.

TinyML is a relatively new machine learning paradigm, yet it is producing remarkably astounding results. Several recent examples include audio, visual, and sensor fusion applications, such as voice and facial recognition, voice commands, and natural language processing.

Looking further into the future, we imagine a world where we have a tiny battery-powered image sensor that I could program to look out for things like particular crop pests or weeds and send an alert when one was spotted. These could be scattered around fields and guide interventions like weeding or pesticides in a much more environmentally friendly way.

The Future of TinyML is Bright

We can conjure up a thousand other products. It feels a lot like being a kid in the Eighties when the first home computers emerged. None of us had any idea what they would become, and most people at the time used them for games or storing address books, but there were so many possibilities. A new world emerged out of those burgeoning systems. So we challenge you to invent the future. Come on and embrace the future with us by learning all we have to teach you about tinyML!