AI Sustainability

The field of artificial intelligence (AI) and machine learning (ML) is growing rapidly and with it the demand for computing power. This is great news for businesses and consumers alike, as it means more powerful and efficient applications are becoming available all the time.

However, this increased demand for computing power comes at an environmental cost. Computing power requires electrical energy, and this energy must be obtained through some means. This energy could come from a variety of sources: wind, solar, geothermal, wave, fossil fuels, and other sources. For both consumers and businesses, this is an upstream process whose specifics are often largely unknown, much like the plumbing or heating systems in an individual's home.

The idea of sustainable energy focuses on the use of energy sources that are, in principle, infinitely renewable (i.e., they can be "sustained" in perpetuity). Because of the nuance involved in energy production, sustainability is not a binary characteristic. For example, the sun has a finite lifetime, albeit very long, which effectively makes it a renewable energy source from a human perspective. On the other hand, coal, oil, and natural gas all exist in finite quantities, although these quantities might differ substantially. Ensuring that the energy used in an AI or ML application is sustainable energy is difficult but is becoming an increasingly prominent issue.

This issue is most important for data centers, which utilize vast numbers of GPUs and other devices in large storage facilities. These types of computers generate significant heat and consume enormous amounts of energy to run their powerful processors; this means that the energy used to power AI and ML is quickly becoming a major contributor to climate change. For instance, in the Continuous Training stage, we discussed how training a single AI model for natural language processing can emit as much carbon as five cars!

So what can be done about this? For the upstream process of energy production, we can help to promote energy governance from a sustainable perspective and also hold companies accountable for their choices of energy. In terms of the models themselves, we aim for two goals: efficiency and minimalism.

Efficiency can be achieved by developing more efficient computing technologies. The best way to characterize efficiency is using a metric like performance per Watt, which can be measured on most architectures using benchmarks such as LINPACK. For a full data center, power usage effectiveness is often used instead as a more holistic metric. Current architectures are already very efficient, but nowhere near the limit of computational power efficiency, known as the Landauer limit. Additionally, the current onus is on the development of more powerful architectures, not power-efficient ones. As a result, our demand for computing using scale-out resources is putting increasing pressure on energy resources, further exacerbating existing environmental issues.

Minimalism can be achieved by optimizing our algorithms so that they require less processing power. This can entail various things, such as reusing models (i.e., transfer learning) so that redundant computations are not performed, and using appropriately sized models, datasets, and architectures for the application, along with the optimal compilation flags and perhaps parallelization. As you might have already guessed, minimalism is really where TinyML comes to life.

One more issue of sustainability we have not touched upon yet is our devices themselves. Every device that we use had to be constructed before we acquired or purchased it. This includes other finite resources such as rare metals, which should not be overlooked from a sustainability perspective. These materials have to be mined, transported, processed, and also deconstructed once their useful life has expired. All of these processes require energy and other additional resources. Thus, in the context of minimalism, we must strive to limit device consumption. Given the huge increase in demand for internet-connected microcontroller devices, this may be the most important challenge to be cognizant of in terms of the sustainability of TinyML.

In the end, the most sustainable option of all is to limit energy usage overall to a minimum. It is up to us as consumers and developers to make sure that AI and ML remain sustainable. We must work together to find ways to reduce energy consumption without sacrificing performance or usability. With a little effort, we can create a future where artificial intelligence and machine learning are both efficient and environmentally friendly.