Beyond connectivity – TinyML

The promise of machine learning's impact on all of society and our lives is a promise of its ubiquity. Pervasive deployment of these learned machines presents a wide range of challenges to go with it. Some of these challenges are of a technical nature, and some are of a social, ethical, and legal one.

On the purely technical side one finds problems relating to the hardware itself. One needs to balance the cost and the computational power of the microcontrollers involved, as well as find ways to supply them with energy and other infrastructural needs, in addition to providing maintenance (or disposability) in a sustainable way. On the more societal side, one sees issues of privacy and data management. And of course, the ever-underappreciated security aspect straddles the two.

The research domain of "Tiny ML" may, very broadly speaking, be said to be concerned with these kinds of issues, with the name referring to the common theme that ubiquitous machine learning devices necessarily must be "tiny" from both a computational, an energy-requirement, and a cost and size perspective. In Telenor Research's TinyML project, we approach this along a wide spectrum that is roughly defined by its two extremes:

The extreme "edge" side: The tiny device should be entirely self-sufficient, extremely cheap, and require no infrastructure other than a very simple energy supply. Such "fire and forget" devices cannot rely on any kind of connectivity, and must perform their machine learning task on their own and in the field, before heavily summarized small-sized results are collected, e.g. through low-power potentially unreliable IoT telecommunication links.

The extreme "cloud" side: The tiny device does nothing on its own, but relies on widespread low-latency, high-bandwidth 4G/5G coverage to offload all its computations to powerful cloud infrastructure. The device is little more than a "streaming sensor", with the actual machine learning happening on more ordinary computing devices in centralized data centers.

Each extreme comes with its own set of advantages and disadvantages. For example, the "extreme edge" is inherently privacy-friendly; a (myriad of) camera(s) in a public space is of no concern if the photos captured are stored as a single number (i.e. the output of the device's machine learning algorithm). It does, however, suffer from extremely constrained computing resources and energy budgets. Conversely, the "extreme cloud" provides the tiny device with all the compute resources of the world's data centres, but at the cost of complicated infrastructure requirements and privacy and security constraints.

These extremes, and the very rich area of tradeoffs in-between — where for example the lines between "on-device" and "off-device" computation are blurred (think e.g. of deep neural networks that run in part on the tiny microcontrollers and in part in the cloud) — will likely be of great importance as Telenor seeks to find its space in the machine learning and artificial intelligence-fueled future.

Contact person

Gard Spreemann, gard.spreemann@telenor.