nanoML: Pushing the Limits of Edge AI with Weightless Neural Networks

Mainstream artificial neural network models, such as Deep Neural Networks (DNNs) are computation-heavy and energy-hungry. Weightless Neural Networks (WNNs) are natively built with RAM-based neurons and represent an entirely distinct type of neural network computing compared to DNNs. WNNs are extremely low-latency, low-energy, and suitable for efficient, accurate, edge inference. The WNN approach derives an implicit inspiration from the decoding process observed in the dendritic trees of biological neurons, making neurons based on Random Access Memories (RAMs) and/or Lookup Tables (LUTs) ready-to-deploy neuromorphic digital circuits. This talk will describe the state of the art of Weightless Neural Networks, and their applications for edge inferencing, especially medical devices.