Deliver the Intelligence at the Edge via On-Device Training of Deep Learning Models on Microcontrollers

Rawan M. A. Nawaiseh[†], Fabrizio De Vita[‡], Dario Bruneo[‡]

[†]*SmartMe.io s.r.l. Messina, Italy*, rawan@smartme.io
[‡]*Department of Engineering, University of Messina, Italy*, {fdevita, dbruneo}@unime.it

Recent advancements in Artificial Intelligence (AI) together with the pervasive presence of Edge devices, represent two triggers for the spreading of Intelligent Cyber Physical Systems (ICPSs) as solution where smart services can run the computing as well as the reasoning phase, without the need of external interactions. Nowadays, with the increase of devices computational power, we are able to execute machine and deep learning models event into low cost hardware such as Micro Controller Units (MCUs) board, giving the birth to the so called Tiny Machine Learning (TinyML). The goal of this paradigm is to shift the intelligence to the Edge for the realization of tailored smart solutions. Unfortunately the hardware constraints and limitations of these devices affect the complexity of the tasks that can be accomplished, for this reason TinyML should not be considered as an alternative to general purpose AI applications, but as a complementary technology.

Until recently, hybrid frameworks have been proposed to address this problem by offloading the more resource demanding tasks to the Cloud (e.g., the training of deep learning models), and performing the inference at the Edge in order to reduce the overall system response time. However, when working in application scenarios where security, latency, and connection stability are crucial requirements, it is clear that this approach is no longer effective. This becomes even more evident especially in those dynamic environments where new data patters can emerge very frequently, thus requiring a constant model training.

In the context of time-series data analysis, most of the Recurrent Neural Networks (RNNs) models such as: Long Short Term Memories (LSTMs) and Gated Recurrent Units (GRUs) are characterized by a large number of trainable parameters which make them unsuitable to be run on a MCU. The use of Reservoir Computing (RC) framework can mitigate this problem by condensing the power and features of RNNs and putting them into a sparse recurrent structure called *reservoir* which exhibits a reduced number of parameters. Echo State Networks (ESNs) are the most popular and representative example of RC models, and consist in a family of architectures where the most part of weights is kept fixed to produce low complexity models with a small memory footprint and very fast training and inference times.

In this work, we propose an anomaly detection algorithm that performs the on-device training of an ESN directly on the MCUs of the STM32 family. To evaluate the effectiveness of our approach, we created a testbed and developed an application that analyzes the vibrations signals recorded by a tri-axial accelerometer, and we used this information to monitor the "health" state of a DC motor. The results obtained from the experimentation demonstrate the feasibility of our solution that enables the training of deep learning models even in devices with limited hardware resources. Moreover, the proposed on-device algorithm is able to effectively detect the occurrence of abnormal conditions inside the testbed and to recognize (almost in real-time) the emergence of new anomalous patterns.