Green Machine Learning for the IoT (GEMINI): a position paper

Luca Lazzaroni[1[0000-0001-8092-5473]], Ali Dabbous[1[0009-0004-8978-4979]], Matteo Nardello[2[0000-0003-3126-1177]], Muath Abu Lebdeh[2[0000-0002-8610-3064]], Riccardo Berta[1[0000-0003-1937-3969]], Kasim Sinan Yildirim[2 [0000-0002-9528-6923]], Francesco Bellotti[1], and Davide Brunelli[2[0000-0001-5110-6823]]

[1] Department of Naval, Electrical, Electronics and Telecommunications Engineering, University of Genoa, 16145 Genoa, Italy
[2] Department of Industrial Engineering, University of Trento, 38122 Trento, Italy
luca.lazzaroni@edu.unige.it

Abstract. The proliferation of IoT devices has significantly increased global energy consumption and carbon footprint due to the reliance on computationally intensive Machine Learning (ML) techniques. Traditionally implemented on high-end embedded devices or exploiting cloud computing, these applications further exacerbate environmental impacts. As IoT technology advances, there's a shift towards smaller, low-power devices, necessitating decentralized computation. This shift has given rise to Tiny Machine Learning (TinyML), enabling ML tasks on edge devices. However, to fully exploit TinyML, integrated frameworks are essential for efficiently developing and deploying smart applications, especially on devices operating intermittently on harvested energy. The Green Machine Learning for the IoT (GEMINI) project aims to address these challenges by creating a framework for sustainable TinyML applications on battery-powered and batteryless energy-harvesting edge devices. GEMINI targets shifting computational workloads from the cloud to the edge, optimizing latency, bandwidth, energy use, and privacy. By integrating energy harvesting methods, zero-power communication protocols, and intermittent computing strategies, GEMINI seeks to ensure robust ML inference under intermittent power conditions. The project aims to include efficient data collection, ML model generation, and online learning directly on edge devices, reducing reliance on constant cloud connectivity. The open-source framework intends to offer a comprehensive toolkit for developing sustainable IoT solutions.

Keywords: Batteryless Device, Energy Harvesting, Intermitting Computing, Internet of Things, Machine Learning, Online Learning, Online Sensing.

1 Introduction

The Internet of Things (IoT) today consists of billions of embedded devices equipped with sensors that perform computations and communicate wirelessly. These devices typically rely on Machine Learning (ML) techniques to enable smart services and applications, enhancing our daily lives by supporting various tasks intelligently. However,

since ML tasks require significant computational power, smart applications typically rely on high-end embedded devices or cloud computing, significantly increasing global energy consumption and the carbon footprint associated with these applications.

As IoT technology continues to advance and the number of IoT devices continues to increase, there is a snowballing shift towards the use of smaller, more resource-constrained, low-power IoT devices. This has led to a growing trend of decentralization of computation, pushing it closer to the edge of the network. This demand has given rise to Tiny Machine Learning (TinyML), a paradigm that enables the execution of ML tasks on small, resource-constrained edge devices. TinyML holds the promise of enabling innovative and environmentally friendly applications and extending the reach of ML to a broader range of IoT devices, including those that operate without batteries by harvesting energy from alternative sources.

To fully realize the potential of TinyML, integrated frameworks are needed to support the efficient development and deployment of smart applications on low-end IoT edge devices. These frameworks must seamlessly combine new computational techniques to execute ML models energy-efficiently, especially on batteryless devices that may operate intermittently due to limited energy availability. Currently, such comprehensive solutions are lacking, hindering the development and deployment of green TinyML applications.

The Green Machine Learning for the IoT (GEMINI) project aims to fill this gap by designing and implementing a novel framework that supports the development of green and fully sustainable TinyML applications distributed across an edge-to-cloud continuum (Fig. 1). The GEMINI framework targets batteryless embedded devices, including those without network connectivity. The project's main objectives are:

1. Facilitate developers in designing TinyML systems that shift computation from the cloud to the edge. This shift offers numerous benefits, including reduced response latency, lower bandwidth usage, decreased energy consumption, a reduced carbon footprint, and enhanced security and privacy.
2. Support the construction of unsupervised models on energy-harvesting, batteryless edge devices, which is crucial for scenarios where edge-to-cloud connectivity is impractical or costly.
3. Enable TinyML models to execute intermittently, despite frequent power failures caused by the sporadic nature of ambient energy.
4. Allow models to evolve and improve during runtime on microcontroller (MCU) platforms by analyzing new data during inference, using an online learning approach.
5. Address the communication phase, enabling the transmission of inference outcomes from the edge (e.g., back to the cloud or to a local base station) using efficient low-power communication techniques. For batteryless edge devices, GEMINI will bypass the cloud model development phase, offering solutions to train and create unsupervised lightweight models on-device.

The project seeks to push the boundaries of TinyML to achieve zero-power ML by providing the following key components: **1)** Efficient data collection; **2)** ML model generation, compression, and optimization for memory-constrained IoT devices; **3)**

Zero-power deployment of ML models using novel communication protocols that exploit backscatter and visible light communication; **4)** Intermittent execution and acceleration of trained ML models on batteryless edge devices, relying solely on energy harvesting (EH). Furthermore, GEMINI aims to significantly reduce the carbon footprint of modern ML-based IoT applications, facilitate knowledge transfer for industrial benefits, and create a new paradigm in the IoT landscape.

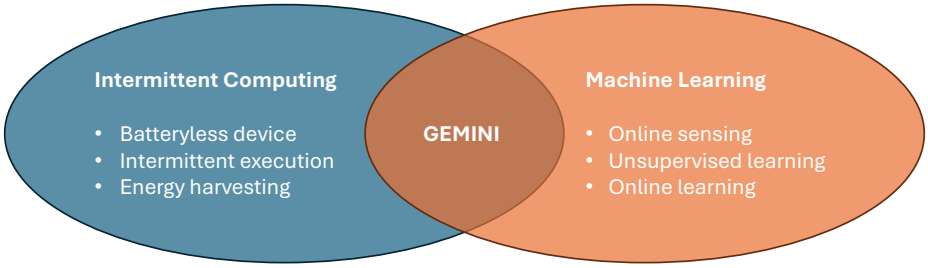

Fig. 1. The GEMINI concept.

2 Related Work

2.1 Intermitting computing

Batteryless devices, which rely on harvesting ambient energy, often face power and energy discontinuities, causing them to operate intermittently [1]. These intermittent operations lead to significant challenges [2]. The most critical consequence of such power failures is the loss of computational state, such as CPU registers and memory contents, which prevents forward computational progress and leads to memory inconsistencies. Traditional software systems, designed for continuously powered computers, fail under intermittent operation. To mitigate this issue, researchers have proposed checkpointing the computational state [3]. When the computational state progresses to a checkpoint, the application saves the registers and volatile memory contents to non-volatile memory. After a power failure, the last checkpoint serves as a backup, allowing computation to resume from a consistent state. Another approach exploits task-based programming models, which offer an efficient alternative to checkpointing but require complex code transformations [4].

2.2 ML on batteryless devices

Recent studies have explored the intermittent execution of data-intensive ML tasks on batteryless edge devices [5–7]. Among these, SONIC [5] and Neuro.ZERO [7] exploit the parallelism provided by the Low Energy Accelerator (LEA) in TI MSP430FR MCUs, which supports vector-based signal processing. Similarly, MANIC [5] is an efficient ultra-low-power vector dataflow architecture. However, both LEA and MANIC

lose their computational state upon power failures, necessitating repeated hardware reconfiguration and data transfer between volatile and non-volatile memory. Finally, TICS [8] offers predictable checkpoint sizes for classification activities. All these studies need to employ software-based intermittent computing techniques, such as checkpoints or task-based models, to ensure resilience to power failures.

2.3 Zero-Power communication

An important aspect to be taken into consideration when deploying on ultra-low-power devices is unexpected power failures, which pose significant challenges to wireless communication, as nodes may become unsynchronized and cease operation during data transmission. Current research has focused on zero-power wireless communication. Backscatter communication is a key technology for zero-power wireless communication in power-constrained, batteryless devices [9]. Most existing backscatter networks perform only reader-to-tag communication, with a single channel between a dedicated master and tags. Recent advancements have enabled communication among batteryless devices (tag-to-tag), allowing each node to function as a transceiver, capable of both encoding and decoding [10]. Furthermore, backscatter communication can be integrated with radiofrequency (RF) EH systems [11], where the RF carrier is used both for backscattering information and for collecting a small amount of energy. Visible Light Communication (VLC) [12, 13] also provides low-power characteristics for the downlink channel between batteryless devices and a central node. Despite these advancements, the reported works require the devices in use to operate continuously during data packet communication. However, batteryless devices function intermittently, leading to numerous failed data transmissions. Thus, a communication stack and protocols that enable batteryless networks are essential for zero-power ML, which requires data collection and model deployment.

3 Methodology

Implementing Deep Neural Networks (DNNs) on resource-constrained embedded IoT devices presents significant challenges, including limited computation, memory, adaptability, and energy [14]. These devices typically have fewer computational units and operate at lower CPU frequencies (1-80 MHz), resulting in long execution times for DNN inference. Additionally, their small memory capacities (1-128 KB) impose stringent size constraints on models, which are often trained only offline on powerful servers. Also, this offline learning approach typically does not include the possibility of learning at runtime, which, instead, could increase the model's accuracy. The most critical challenge, however, is energy as most IoT devices rely on batteries that can deplete within days or even hours.

GEMINI aims to address these challenges through the development of an intelligent intermittent platform following a hardware-software co-design approach. From a hardware perspective, by leveraging EH methods that efficiently scavenge energy from the ambient environment, the GEMINI platform will sustain computations even with an

intermittent and low power supply and enable near-zero energy communication with other IoT devices through transient communication. From a software perspective, an ML optimizer framework will be designed to operate with the models generated on the batteryless nodes. While ML execution is inherently time and energy-consuming, the use of EH will ensure that power outages will only delay the output without affecting the inference results.

GEMINI's framework will be completely open-source, aiming to support researchers in designing and deploying ML-based IoT solutions on edge devices. The framework relies on two main pillars: **1)** Cloud-side, responsible for data collection, model training, and optimization for resource-constrained platforms; **2)** Node-side, which will execute a platform-independent runtime on resources-constrained devices for ML inference, based on the capabilities of the underlying hardware. In this two-headed context, GEMINI will follow a three steps workflow:

1. Zero-Power data collection: Edge devices collect field data to build a suitable dataset for model generation [15]. Data are timestamped, logged, compressed, and sent to a central node using low-power, secure communication mechanisms, preserving the privacy of the collected information [16]. Batteryless edge devices, which experience frequent power failures and operate intermittently, will use low data-rate, zero-power communication techniques such as backscatter and visible light communication for data transmission. Synchronization, security, and energy-aware data collection protocols will also be explored.
2. Tiny model generation: Users select the ML task and algorithm through GEMINI on the cloud, utilizing libraries like PyTorch and TensorFlow. If data are not pre labeled, unsupervised classification using clustering algorithms is performed. Hyperparameter tuning is conducted using grid or random search methods. Models are optimized for specific hardware constraints through techniques like pruning and quantization and stored for deployment using low-power communication methods (e.g., [17, 18]). Different model versions may be created to balance accuracy and energy consumption, also exploring Neural Architecture Search (NAS) techniques.
3. Zero-Power TinyML inference and online learning: The GEMINI edge runtime will support low-power MCUs and accelerators, such as the LEA vector accelerator and convolutional neural network (CNN) accelerators. The runtime aims to execute ML inference resiliently against power failures and support on-device online unsupervised learning when communication is infeasible.

From the user perspective, the coordination and execution of the three steps will be facilitated using an ultra-low-power Voice User Interface (VUI) (e.g., [19]), making any user able to control devices with voice by giving specific commands, thus increasing versatility and personalization. GEMINI will be validated across various IoT contexts, such as smart homes and automated vehicle environments, to assess performance limits, power consumption, and development workflow efficiency. Applications will include natural language processing for voice assistants, object detection through surveillance cameras, and fault detection in automated vehicles. Continuous monitoring applications, like air quality learning on energy-harvesting batteryless devices, will also be explored.

GEMINI's expected outcomes include:

1. An open-source edge-cloud framework for data collection, TinyML model building, and optimization.
2. A platform-independent runtime system for executing TinyML models on edge devices.
3. A hardware prototype demonstrating transient communication, intermittent inference, and TinyML functionalities.
4. Prototype applications in smart home and automated vehicle scenarios to validate the framework's effectiveness.

An example of possible improvements achievable by exploiting intermittency execution on the edge [20] is presented in Fig. 2 that shows how it is possible to increase the number of inferences executed by exploiting an intermittent toolchain that implements energy monitoring mechanism, i.e., capacitor energy monitoring policy (CMPCh). Thanks to energy monitoring mechanism and energy-aware-checkpointing it is possible to increase the number of inferences completed even when the baseline approach fails due to the extremely low energy input.

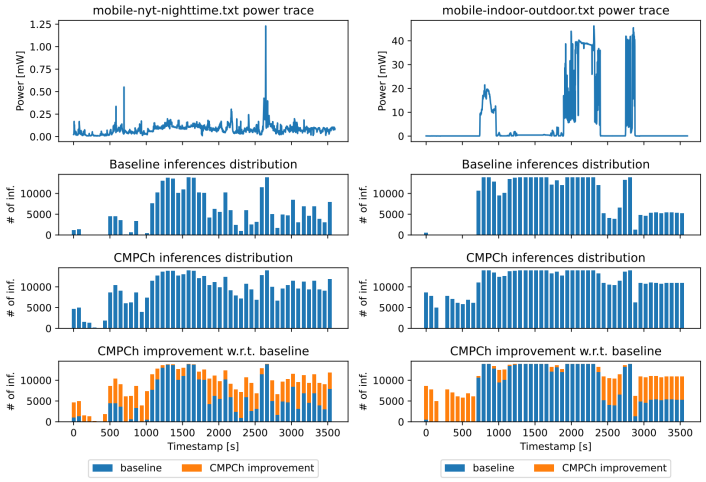

Fig. 2. Throughput comparison with and without an intermittent-enabled toolchain [20].

4 Conclusion

The GEMINI project targets a significant advancement in the field of IoT and TinyML by addressing the critical challenges of computation, memory, adaptability, and energy efficiency. By leveraging EH techniques and developing a robust hardware-software co-design approach, GEMINI promises to enable the deployment of intelligent, batteryless IoT devices capable of sustainable operation. The project's innovative method-

ologies (including intermittent computing, transient communication, and online learning) look toward zero-power ML applications. The open-source nature of the GEMINI framework will foster collaboration, facilitating the development of a wide range of IoT applications across various domains. Ultimately, GEMINI aims to reduce the carbon footprint of IoT applications, promote the adoption of green technologies, and contribute to the broader goal of sustainable development in the tech industry.

Acknowledgements

This work has been supported by the GEMINI "Green Machine Learning for the IoT" national research project, funded by the MUR under the PRIN 2022 program (Contract 20223M4HZ4).

References

1. Doglioni, M., Nardello, M., Brunelli, D.: Plant Microbial Fuel Cells: Energy Sources and Biosensors for battery-Free Smart Agriculture. IEEE Transactions on AgriFood Electronics. 1–11 (2024). https://doi.org/10.1109/TAFE.2024.3417644.
2. Rodriguez Arreola, A., Balsamo, D., Das, A.K., Weddell, A.S., Brunelli, D., Al-Hashimi, B.M., Merrett, G.V.: Approaches to Transient Computing for Energy Harvesting Systems: A Quantitative Evaluation. In: Proceedings of the 3rd International Workshop on Energy Harvesting & Energy Neutral Sensing Systems. pp. 3–8. Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2820645.2820652.
3. Jayakumar, H., Raha, A., Lee, W.S., Raghunathan, V.: QuickRecall: A HW/SW Approach for Computing across Power Cycles in Transiently Powered Computers. J. Emerg. Technol. Comput. Syst. 12, 8:1-8:19 (2015). https://doi.org/10.1145/2700249.
4. Mülayim, M., Goknil, A., Yıldırım, K.S.: Taskify: An Integrated Development Environment to Develop and Debug Intermittent Software for the Batteryless Internet of Things. In: 2020 16th International Conference on Distributed Computing in Sensor Systems (DCOSS). pp. 348–355 (2020). https://doi.org/10.1109/DCOSS49796.2020.00062.
5. Gobieski, G., Lucia, B., Beckmann, N.: Intelligence Beyond the Edge: Inference on Intermittent Embedded Systems. In: Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems. pp. 199–213. Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3297858.3304011.
6. Gobieski, G., Nagi, A., Serafin, N., Isgenc, M.M., Beckmann, N., Lucia, B.: MANIC: A Vector-Dataflow Architecture for Ultra-Low-Power Embedded Systems. In: Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture. pp. 670–684. Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3352460.3358277.
7. Lee, S., Nirjon, S.: Neuro.ZERO: a zero-energy neural network accelerator for embedded sensing and inference systems. In: Proceedings of the 17th Conference on Embedded Networked Sensor Systems. pp. 138–152. Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3356250.3360030.

8. Kortbeek, V., Yildirim, K.S., Bakar, A., Sorber, J., Hester, J., Pawełczak, P.: Time-sensitive Intermittent Computing Meets Legacy Software. In: Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems. pp. 85–99. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3373376.3378476.
9. Galisteo, A., Varshney, A., Giustiniano, D.: Two to tango: hybrid light and backscatter networks for next billion devices. In: Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services. pp. 80–93. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3386901.3388918.
10. Majid, A.Y., Jansen, M., Delgado, G.O., Yildirim, K.S., Pawełłzak, P.: Multi-hop Backscatter Tag-to-Tag Networks. In: IEEE INFOCOM 2019 - IEEE Conference on Computer Communications. pp. 721–729 (2019). https://doi.org/10.1109/INFOCOM.2019.8737551.
11. Torrisi, A., Yıldırım, K.S., Brunelli, D.: Reliable Transiently-Powered Communication. IEEE Sensors Journal. 22, 9124–9134 (2022). https://doi.org/10.1109/JSEN.2022.3158736.
12. Varshney, A., Soleiman, A., Mottola, L., Voigt, T.: Battery-free Visible Light Sensing. In: Proceedings of the 4th ACM Workshop on Visible Light Communication Systems. pp. 3–8. Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3129881.3129890.
13. Rehman, S.U., Ullah, S., Chong, P.H.J., Yongchareon, S., Komosny, D.: Visible Light Communication: A System Perspective—Overview and Challenges. Sensors. 19, 1153 (2019). https://doi.org/10.3390/s19051153.
14. Dabbous, A., Berta, R., Fresta, M., Ballout, H., Lazzaroni, L., Bellotti, F.: Bringing Intelligence to the Edge for Structural Health Monitoring: The Case Study of the Z24 Bridge. IEEE Open Journal of the Industrial Electronics Society. 5, 781–794 (2024). https://doi.org/10.1109/OJIES.2024.3434341.
15. Fresta, M., Bellotti, F., Capello, A., Dabbous, A., Lazzaroni, L., Ansovini, F., Berta, R.: End-to-End Dataset Collection System for Sport Activities. Electronics. 13, 1286 (2024). https://doi.org/10.3390/electronics13071286.
16. Berta, R., Bellotti, F., De Gloria, A., Lazzaroni, L.: Assessing Versatility of a Generic End-to-End Platform for IoT Ecosystem Applications. Sensors. 22, 713 (2022). https://doi.org/10.3390/s22030713.
17. Sakr, F., Berta, R., Doyle, J., Capello, A., Dabbous, A., Lazzaroni, L., Bellotti, F.: CBin-NN: An Inference Engine for Binarized Neural Networks. Electronics. 13, 1624 (2024). https://doi.org/10.3390/electronics13091624.
18. Berta, R., Dabbous, A., Lazzaroni, L., Pau, D., Bellotti, F.: Developing a TinyML Image Classifier in a Hour. IEEE Open Journal of the Industrial Electronics Society. 1–15 (2024). https://doi.org/10.1109/OJIES.2024.3451959.
19. Lazzaroni, L., Bellotti, F., Berta, R.: An embedded end-to-end voice assistant. Engineering Applications of Artificial Intelligence. 136, 108998 (2024). https://doi.org/10.1016/j.engappai.2024.108998.
20. Caronti, L., Akhunov, K., Nardello, M., Yıldırım, K.S., Brunelli, D.: Fine-grained Hardware Acceleration for Efficient Batteryless Intermittent Inference on the Edge. ACM Trans. Embed. Comput. Syst. 22, 82:1-82:19 (2023). https://doi.org/10.1145/3608475.