Poster: Towards Battery-Free Machine Learning Inference and Model Personalization on MCUs

Yushan Huang
Imperial College London
London, UK
yushan.huang21@imperial.ac.uk

Hamed Haddadi
Imperial College London
London, UK
h.haddadi@imperial.ac.uk

ABSTRACT

Machine learning (ML) is moving towards edge devices. However, ML models with high computational demands and energy consumption pose challenges for ML inference in resource-constrained environments, such as the deep sea. To address these challenges, we propose a battery-free ML inference and model personalization pipeline for microcontroller units (MCUs). As an example, we performed fish image recognition in the ocean. We evaluated and compared the accuracy, runtime, power, and energy consumption of the model before and after optimization. The results demonstrate that, our pipeline can achieve 97.78% accuracy with 483.82 **KB** Flash, 70.32 **KB** RAM, 118 **ms** runtime, 4.83 **mW** power, and 0.57 **mJ** energy consumption on MCUs, reducing by 64.17%, 12.31%, 52.42%, 63.74%, and 82.67%, compared to the baseline. The results indicate the feasibility of battery-free ML inference on MCUs.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; • **Computer systems organization** → **Embedded systems**.

KEYWORDS

Edge Computing, IoT, TinyML, Resource-constrained

ACM Reference Format:
Yushan Huang and Hamed Haddadi. 2023. Poster: Towards Battery-Free Machine Learning Inference and Model Personalization on MCUs. In *The 21st Annual International Conference on Mobile Systems, Applications and Services (MobiSys '23), June 18–22, 2023, Helsinki, Finland.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3581791.3597371

1 INTRODUCTION

Machine learning (ML) models have become ubiquitous in solving diverse problems. However, these models often demand high memory, computation power, and energy requirements, posing challenges for ML deployment on resource-constrained edge devices. The edge offers several advantages such as reduced response latency, better bandwidth utilization, and improved security and privacy expectations. Therefore, there is a pressing need to develop optimized lightweight ML models for deployment on the edge.

MobiSys '23, June 18–22, 2023, Helsinki, Finland

ACM ISBN 979-8-4007-0110-8/23/06.
https://doi.org/10.1145/3581791.3597371

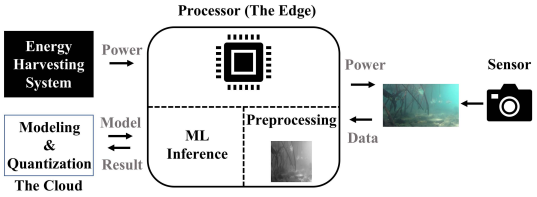

Figure 1: Design of battery-free inference on MCUs

To address this problem, researchers have explored model compression techniques to compress the model for the edge [1]. However, these technologies typically assume that the edge has sufficient memory, computing power, and power supply, which can be challenging to achieve in extreme environments such as deep seas, and remote areas. Some studies have deployed traditional models such as SVM on MCUs [2], but they require manual feature extraction and may not perform well on high-dimensional data. Additionally, power and energy consumption are often overlooked in edge-based ML deployment. Recent advancements in battery-free sensing technology have allowed for innovative applications [3]. These techniques have made long-term MCUs use possible without the need for power supplies or batteries. this study aims to examine the feasibility of achieving battery-free ML inference and model personalization on MCUs for extreme environments.

In our previous work, we have achieved battery-free inference for sound signal classification [4]. However, we did not consider optimizing the model and energy consumption. The primary contributions of this paper are as follows:

(1) This pipeline is an end-to-end solution designed for the efficient design, deployment, energy-optimizing, and execution of ML models on resource-constrained MCUs.

(2) We optimized the Proxyless neural architecture search (ProxylessNAS) [5] model, resulting in a reduction of the model size from 1350.25 **KB** Flash and 80.20 **KB** RAM to 483.82 **KB** Flash and 70.32 **KB** RAM, with a slight loss of approximately 0.1% in accuracy.

(3) We compared the runtime, power, and energy consumption of the model before and after optimization by Monsoon Power Monitor [6]. Compared with the unoptimized model, we achieved 97.78% accuracy with a runtime of 118 **ms**, power of 4.83 **mW**, and energy consumption of 0.57 **mJ**, which reduced by 52.42%, 63.74%, and 82.67%, respectively.

2 PIPELINE

The pipeline mainly has four components: preprocessing, modeling, quantization, and transplantation and inference, as shown in Fig. 1.

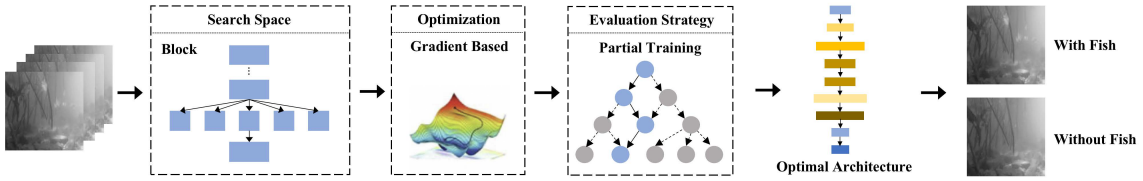

Figure 2: The main components of ProxylessNAS.

Preprocessing We conducted a study on fish recognition in the ocean using DeepFish dataset [7]. DeepFish is a dataset based on field records of fish habitats and customized for the analysis of fish in underwater marine environments. The dataset comprises approximately 40,000 high-resolution 1,920×1,080 data points. As Fig. 1 (a) shows, the original images with three RGB channels at a resolution of 1,920×1,080 are reconstructed into single RGB channel images at a resolution of 32×32.

Modeling. We optimized the ProxylessNAS [5] to search the approximate model. ProxylessNAS aims to directly search for architectures for target tasks, and optimizes model weights and architecture parameters alternatively using gradient-based methods. ProxylessNAS introduces binary gates that binarize the architecture parameters of an overparameterized network. This enables only one path to load at runtime, reducing memory consumption by not loading the entire overparameterized network to update model weights. The main components of ProxylessNas are shown in Fig. 2.

Quantization. To reduce the model's size, computation requirements, power, and energy consumption, we utilize static quantization to quantize the original model. Static quantization is an optimization technique for neural network models, which converts the parameters and activations from floating-point to integer representations while preserving the model's accuracy. The process involves data collection, quantization range determination, quantization conversion, and dequantization. Static quantization significantly reduces model size, speeds up inference, and lowers power and energy consumption while maintaining model accuracy.

Transplantation. We utilize X-CUBE-AI to transplant the model and inference process into *.h* and *.c* files. X-CUBE-AI is a software package that helps developers integrate models into embedded applications on MCUs. It includes tools such as a neural network model converter and inference engine.

3 EVALUATION

We trained the lightweight ProxylessNAS model by TensorFlow, transferred it to TFLite, and applied static quantization to reduce the model size, computation requirement, power, and energy consumption. We evaluated and compared the offline accuracy, and conducted 10 repeated experiments on the original TensorFlow model, original TFLite model, and optimized TFLite model. The accuracies of these models are 97.88±1.12%, 97.88±1.12%, and 97.78±1.08%. The quantized TFLite Model only loses approximately 0.1% accuracy.

We transplanted the TFLite models to the STM32L4R5 development kit, and utilized the Monsoon power monitor to measure power and energy. Power was calculated as the product of the current and voltage when the board was connected to the power source. The board was supplied power at 1.9**V**. The power consumption results are presented in Table. 1. Previous research has shown that underwater acoustic and ultrasonic signals can generate a few **mW** [3], suggesting that our pipeline can run solely on harvested energy.

Table 1: Evaluation of the power performance on MCUs

	Flash (**KB**)	RAM (**KB**)	Power (**mW**)	Time (**ms**)	Energy (**mJ**)
Original Model	1350.25	80.20	13.32	248	3.29
Optimized Model	483.82	70.32	4.83	118	0.57

4 CONCLUSION AND FUTURE WORK

This paper introduces an energy-optimized ML deployment pipeline for resource-constrained MCUs. Compared to the unoptimized model, we achieved an average accuracy of 97.78% with 483.82 **KB** Flash, 70.32 **KB** RAM, 118 **ms** inference time, 4.83 **mW** power, and 0.57 **mJ** energy consumption, which reduced by 64.17%, 12.31%, 52.42%, 63.74%, and 82.67%, respectively. The results indicate the viability of battery-free ML on MCUs, with the potential to harvest energy from certain devices. In the future, we plan to further optimize energy consumption and improve the personalization of the model to address the data heterogeneity problem. We will also test the pipeline on various models and tasks to assess its scalability.

REFERENCES

[1] Wei Niu, Xiaolong Ma, Sheng Lin, Shihao Wang, Xuehai Qian, Xue Lin, Yanzhi Wang, and Bin Ren. Patdnn: Achieving real-time dnn execution on mobile devices with pattern-based weight pruning. In *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*, pages 907–922, 2020.

[2] Bharath Sudharsan, John G Breslin, and Muhammad Intizar Ali. Edge2train: A framework to train machine learning models (svms) on resource-constrained iot edge devices. In *Proceedings of the 10th Int'l Conf. on the IoT*, pages 1–8, 2020.

[3] Raffaele Guida, Emrecan Demirors, Neil Dave, and Tommaso Melodia. Underwater ultrasonic wireless power transfer: A battery-less platform for the internet of underwater things. *IEEE Transactions on Mobile Computing*, 21(5):1861–1873, 2020.

[4] Yuchen Zhao, Sayed Saad Afzal, Waleed Akbar, Osvy Rodriguez, Fan Mo, David Boyle, Fadel Adib, and Hamed Haddadi. Towards battery-free machine learning and inference in underwater environments. In *Proceedings of the 23rd Annual International Workshop on Mobile Computing Systems and Applications*, HotMobile '22, page 29–34, New York, NY, USA, 2022. Association for Computing Machinery.

[5] Han Cai, Ligeng Zhu, and Song Han. Proxylessnas: Direct neural architecture search on target task and hardware. *arXiv preprint arXiv:1812.00332*, 2018.

[6] Monsoon solutions inc. https://www.msoon.com/.

[7] Alzayat Saleh, Issam H Laradji, Dmitry A Konovalov, Michael Bradley, David Vazquez, and Marcus Sheaves. A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. *Scientific Reports*, 10(1):14671, 2020.