Special Issue on TinyML

Vijay Janapa Reddi, *Harvard University, Cambridge, MA, 02138, USA*
Boris Murmann, *University of Hawai'i at Mānoa, Honolulu, HI, 96822, USA*

Welcome to *IEEE Micro*'s Special Issue on Tiny Machine Learning (TinyML). This special issue of *IEEE Micro* explores cutting-edge research on TinyML.

ML has transformed numerous fields in recent years, enabled by powerful deep neural networks running on servers with abundant computational resources. However, a quiet revolution has also been unfolding in the domain of embedded systems—ML models designed to run on microcontrollers and other highly resource-constrained devices. This special issue highlights the remarkable progress and potential of TinyML and why it represents a key evolution in making AI ubiquitous.

The concept of TinyML originated about five years ago, as researchers began exploring methods to compress and optimize neural networks to run on low-power hardware. This opened up exciting new possibilities for on-device ML using "tiny" microcontrollers with limited memory and processing capabilities. TinyML broke down the barrier to bringing AI to extreme edge devices like smartphones, wearables, sensors, Internet of Things (IoT) nodes, and more by enabling deep learning locally on the device without relying on connectivity to the cloud. This spurred rapid adoption across industries.

TinyML now powers smart speakers in homes that recognize wake words, mobile apps that identify images or translate languages offline, industrial IoT sensors that detect anomalies, and wearables that monitor health metrics. The applications are endless—from augmented reality/VR, robotics, and autonomous vehicles to agriculture, retail, smart cities, and more.

Key innovations that have enabled this revolution include neural network pruning, quantization, efficient architectures like MobileNets, and new training techniques to create small yet accurate models. For instance, it is now possible to compress large deep learning models by more than 95% without a significant loss of accuracy. Combined with ultralow-power microcontrollers designed specifically for ML, TinyML delivers high accuracy with minimal latency, security, and energy consumption.

Despite these advances, there remain numerous research challenges and opportunities. Further optimizations in model compression, energy-efficient inference, specialized low-power hardware, and robust training techniques have tremendous potential to unlock several new use cases, some of which we have yet to envision. TinyML frameworks that simplify development pipelines will accelerate adoption. Moreover, applications like personalized health monitoring, privacy-centric computing, and bias mitigation can benefit greatly from advances in on-device learning. The opportunities are boundless.

> *THIS SPECIAL ISSUE HIGHLIGHTS THE REMARKABLE PROGRESS AND POTENTIAL OF TinyML AND WHY IT REPRESENTS A KEY EVOLUTION IN MAKING AI UBIQUITOUS.*

The confluence of ML, microelectronics, and embedded systems makes TinyML inherently cross disciplinary. The pace of progress is rapid, and the possibilities are exciting. With TinyML, we envision an intelligent, connected world where even the smallest devices can perceive, reason, and interact automatically. The foundation has been laid, and this is just the beginning—TinyML is and will be a transformative force in shaping the future as AI permeates the physical world.

THE FUTURE OF ML IS TINY AND BRIGHT!

This special issue was created to help spotlight innovative work within the landscape of TinyML today, covering both research innovations and practical deployments. The call for articles invited academic and industry researchers to address issues that span various topics related to TinyML, including datasets, applications,

0272-1732 © 2023 IEEE
Digital Object Identifier 10.1109/MM.2023.3322048
Date of current version 2 November 2023.

algorithms, systems, software, hardware, and evaluation methods. These topics cover the full stack for deploying ML capabilities on resource-constrained edge devices.

We received a total of 22 submissions. The submissions describe innovations in federated learning, neural architecture search, hardware accelerators, power optimization, ML deployments at scale, performance evaluation of TinyML systems, and more. The articles provided insights into the state of the art in the fast-moving field of TinyML across research and practical deployment. Alas, we could only accept a few articles.

We thank all of the authors who submitted articles and the many reviewers who helped with the review process, leading to the careful selection of nine articles for inclusion in the special issue. The accepted works showcase many techniques across diverse TinyML research and development facets.

Of the accepted articles, some present methods to optimize ML models for efficient deployment, including on-device learning from a few examples, in situ distillation to boost tiny model accuracy, and systematic multiobjective fine-tuning leveraging regression and metric learning. Novel architectures and implementations include single-chip neuromorphic processors inspired by dendrites and hardware–software co-design for runtime optimization.

Other articles include strategies for on-device adaptation, techniques to customize models to their deployed environment, and discussions of reducing the gap between training and inference distributions. The articles also include specialized TinyML applications, such as anomaly detection in time series data and metareasoning for multigoal reinforcement learning.

These innovations collectively advance capabilities across on-device training, optimized neural architectures, efficient inference implementations, and cutting-edge applications. They address critical performance, efficiency, and real-world applicability challenges in deploying TinyML solutions.

We hope this curated collection of the latest TinyML research provides a snapshot of the state of the art and a source of inspiration for future work. We are excited by the possibilities opened up as TinyML enables increasingly sophisticated intelligence at the very edge. We thank all contributing authors and researchers for pushing this field forward.

In the following sections, we highlight the contributions of each of the accepted articles. The articles can be grouped into the following main topics.

ENERGY-EFFICIENT ML

Energy-efficient ML is a foundational concern in TinyML research. This core challenge is addressed through four articles. In "Making Machine Learning More Energy Efficient by Bringing It Closer to the Sensor,"[A1] the authors propose a hybrid near-sensor, in-sensor convolutional neural network (CNN) and decision tree approach for motion classification that reduces energy consumption by minimizing data transfer. The combination achieves good accuracy while lowering energy use.

In "A 10.7-μJ/Frame 88% Accuracy CIFAR-10 Single-Chip Neuromorphic Field-Programmable Gate Array Processor Featuring Various Nonlinear Functions of Dendrites in the Human Cerebrum,"[A2] the authors present a spiking CNN neuromorphic architecture implemented on a single field-programmable gate array (FPGA). By mimicking nonlinear dendrite functions and using a line scan design, they achieved an energy efficiency of 10.7 μJ/frame and 88% accuracy on Canadian Institute for Advanced Research, 10 classes (CIFAR-10).

WE HOPE THIS CURATED COLLECTION OF THE LATEST TinyML RESEARCH PROVIDES A SNAPSHOT OF THE STATE OF THE ART AND A SOURCE OF INSPIRATION FOR FUTURE WORK.

In "MetaE2RL: Toward Meta-Reasoning for Energy-Efficient Multigoal Reinforcement Learning With Squeezed-Edge You Only Look Once,"[A3] the authors incorporate metareasoning to switch between reinforcement learning models of different complexities and squeezed edge you only look once (YOLO) for efficient preprocessing. Implemented on drones and Jetson Nano, it improves energy efficiency by 15% and throughput by 21% with a success rate of more than 80%.

Finally, in "Exploring Memory-Oriented Design Optimization of Edge AI Hardware for Extended Reality Applications,"[A4] the authors investigate the benefits of nonvolatile memory in the inference pipeline's memory hierarchy. The authors show that energy savings of 24% can be achieved for activities like hand detection and eye segmentation tasks.

ON-DEVICE CUSTOMIZATION AND ADAPTATION

On-device customization and adaptation of ML models are crucial for enabling effective TinyML applications. This ability to tailor models directly on resource-constrained devices allows for greater flexibility, personalization, and autonomy—all critical capabilities for advanced TinyML systems. The progress made on this

APPENDIX: RELATED ARTICLES

A1. M. Brehler, L. Camphausen, B. Heidebroek, D. Krön, H. Gründer, and S. Camphausen, "Making machine learning more energy efficient by bringing it closer to the sensor," *IEEE Micro*, vol. 43, no. 6, pp. 11–18, Nov./Dec. 2023, doi: 10.1109/MM.2023.3316348.

A2. A. Kosuge, Y.-C. Hsu, R. Sumikawa, M. Hamada, T. Kuroda, and T. Ishikawa, "A 10.7-μJ/frame 88% accuracy CIFAR-10 single-chip neuromorphic field-programmable gate array processor featuring various nonlinear functions of dendrites in the human cerebrum," *IEEE Micro*, vol. 43, no. 6, pp. 19–27, Nov./Dec. 2023, doi: 10.1109/MM.2023.3315676.

A3. M. Navardi, E. Humes, T. Manjunath, and T. Mohsenin, "MetaE2RL: Toward meta-reasoning for energy-efficient multigoal reinforcement learning with squeezed-edge you only look once," *IEEE Micro*, vol. 43, no. 6, pp. 29–39, Nov./Dec. 2023, doi: 10.1109/MM.2023.3318200.

A4. V. Parmar, S. S. Sarwar, Z. Li, H.-H. S. Lee, B. De Salvo, and M. Suri, "Exploring memory-oriented design optimization of edge AI hardware for extended reality applications," *IEEE Micro*, vol. 43, no. 6, pp. 40–49, Nov./Dec. 2023, doi: 10.1109/MM.2023.3321249.

A5. M. Rusci and T. Tuytelaars, "On-device customization of tiny deep learning models for keyword spotting with few examples," *IEEE Micro*, vol. 43, no. 6, pp. 50–57, Nov./Dec. 2023, doi: 10.1109/MM.2023.3311826.

A6. E. S. Pereira, L. S. Marcondes, and J. M. Silva, "On-device tiny machine learning for anomaly detection based on the extreme values theory," *IEEE Micro*, vol. 43, no. 6, pp. 58–65, Nov./Dec. 2023, doi: 10.1109/MM.2023.3316918.

A7. K. Sunaga, M. Kondo, and H. Matsutani, "Addressing the gap between training data and deployed environment by on-device learning," *IEEE Micro*, vol. 43, no. 6, pp. 66–73, Nov./Dec. 2023, doi: 10.1109/MM.2023.3314711.

A8. A. N. Mazumder and T. Mohsenin, "Reg-TuneV2: A hardware-aware and multiobjective regression-based fine-tuning approach for deep neural networks on embedded platforms," *IEEE Micro*, vol. 43, no. 6, pp. 74–83, Nov./Dec. 2023, doi: 10.1109/MM.2023.3316433.

A9. S. Zhang, Y. Fu, S. Wu, J. Dass, H. You, and Y. Lin, "NetDistiller: Empowering tiny deep learning via in situ distillation," *IEEE Micro*, vol. 43, no. 6, pp. 84–92, Nov./Dec. 2023, doi: 10.1109/MM.2023.3324261.

A10. P. Behnam et al., "Hardware–software co-design for real-time latency–accuracy navigation in tiny machine learning applications," *IEEE Micro*, vol. 43, no. 6, pp. 93–101, Nov./Dec. 2023, doi: 10.1109/MM.2023.3317243.

front, as covered by the three articles in this section, marks a step toward more powerful and versatile TinyML solutions.

In "On-Device Customization of Tiny Deep Learning Models for Keyword Spotting With Few Examples,"[A5] the authors show that keyword spotting models can be customized using only 10 examples, achieving good accuracy without extra training—thus enabling easy customization for tiny devices.

In "On-Device Tiny Machine Learning for Anomaly Detection Based on the Extreme Values Theory,"[A6] the authors show how to leverage the Weibull distribution for unsupervised on-device anomaly detection in time series. Their system identifies anomalies efficiently with 99.8% accuracy on synthetic data.

Last but not least, in "Addressing the Gap Between Training Data and Deployed Environment by On-Device Learning,"[A7] the authors retrain ML models on device to adapt to deployed environments. Experiments on vibration data show that this improves accuracy and saves communication costs.

MODEL COMPRESSION AND OPTIMIZATION

Model compression and optimization techniques are essential for deploying capable ML models within the tight resource constraints of TinyML devices. By reducing model size and computational requirements without excessively sacrificing accuracy, the contributions in this area enable more advanced ML applications on extremely limited hardware.

The continuing research advancements on compressing and streamlining ML models are integral

to overcoming the inherent challenges of on-device deployment, making model optimization a key optimization pillar of TinyML progress. To that end, we have two articles on this topic. In "Reg-TuneV2: A Hardware-Aware and Multiobjective Regression-Based Fine-Tuning Approach for Deep Neural Networks on Embedded Platforms,"[A8] the authors show how to compress deep neural networks considering accuracy, power, and latency using regression and metric learning. The method reduces the model size by 14.5× for keyword spotting on FPGA. In "NetDistiller: Empowering Tiny Deep Learning via In Situ Distillation,"[A9] the authors propose a framework for boosting the accuracy of tiny neural networks (NNs) by incorporating the target tiny NN as a student subnetwork within a larger weight-sharing teacher network. Their ideas enable in-situ knowledge transfer from teacher to student during joint training. To do this they have to address several challenges, but their experiments show NetDistiller outperforms state-of-the-art training techniques, improving accuracy by up to 4.5% on ImageNet classification and 1.9% when transferred to object detection.

HARDWARE–SOFTWARE CO-DESIGN

Hardware–software co-design is essential to fully realizing the potential of TinyML systems. More efficient, high-performance, and capable TinyML solutions can be developed by holistically optimizing both the hardware and software aspects in tandem.

In "Hardware–Software Co-design for Real-Time Latency–Accuracy Navigation in Tiny Machine Learning Applications,"[A10] the authors propose SUSHI. This hardware–software co-design caches subgraphs to optimize for latency and accuracy. It improves latency by 32% and accuracy by 1% across neural architectures.

In summary, these articles showcase exciting new techniques like efficient neural architectures, on-device learning, and hardware–software co-design to advance the state of the art in deploying TinyML solutions optimized for the edge. They provide valuable insights into the current challenges and innovations in enabling ML applications on resource-constrained devices. The ideas presented will impact researchers and practitioners aiming to build intelligent systems at the edge.

CONCLUSION

While the articles accepted for this special issue demonstrate important progress, opportunities remain to expand TinyML across applications and advance techniques like on-device learning. Additional innovation around specialized datasets, security protections, continual learning, and low-cost hardware will unlock the full potential of TinyML. The rapid pace of research across the cross-disciplinary intersections of ML, microelectronics, and embedded systems promises many more breakthroughs.

There are many more exciting frontiers to explore as we build an intelligent, perceptive, and autonomous world enabled by TinyML. Thus, while current progress is significant, expanding datasets, securing systems, advancing on-device learning, lowering hardware costs, and robust evaluation represent exciting frontiers for the future of TinyML. The next wave of articles will likely enrich the field across these and many other dimensions.

We invite researchers to push boundaries on what is possible on highly constrained devices. Let us cocreate a future fused with on-device intelligence to enrich our lives. The future of ML is tiny and bright!

VIJAY JANAPA REDDI is an associate professor at Harvard University, Cambridge, MA, 02138, USA, as well as vice president and a founding member of MLCommons (https://mlcommons.org), a nonprofit organization devoted to accelerating machine learning innovation for all. Contact him at vj@eecs.harvard.edu.

BORIS MURMANN is a professor in the Department of Electrical and Computer Engineering, University of Hawai'i at Mānoa, Honolulu, HI, 96822, USA. Contact him at bmurmann@hawaii.edu.