Abstract

In recent years, the Internet of Things (IoT) ecosystem has been increasingly populated by tiny smart devices, typically powered by Microcontroller Units (MCUs), capable of locally sensing and processing user and environment data thanks to Deep Neural Networks (DNNs). However, the severe limitations in computational capabilities and memory size of these ultra-low-power (ULP) platforms hardly limit the size of DNNs deployed in the field, also limiting their generalization capabilities. Therefore, these tiny Artificial Intelligence models, conventionally trained on powerful servers before being deployed frozen on edge devices, often fail when the data encountered in the field differ significantly from the training distribution. To address this *domain shift* issue, the On-Device Learning (ODL) paradigm has emerged as a promising opportunity to enable DNNs to adapt directly in the field, without relying on external resources, promising improved privacy, update latency, and communication cost. This dissertation advances the field of ODL by introducing a full-stack methodology for the efficient fine-tuning of compact DNNs on MCU-powered IoT nodes. Targeting state-of-the-art Parallel Ultra-Low-Power (PULP) RISC-V multi-core MCUs, our deployment strategy is structured on three levels of abstraction. At the firmware level, we introduce PULP-TrainLib, a hardware-optimized library that enables high-performance training with both 32-bit and 16-bit floating-point operations, achieving up to 6.62 MAC/clk on an 8-core MCU thanks to parallel execution, reduced-precision Single-Instruction-Multiple-Data (SIMD) instructions, and properly tuned data reshape operators. On the training algorithm side, we propose structured sparse update schemes that significantly reduce the memory footprint and computational cost of backpropagation, while maintaining accuracy across image classification and monocular depth estimation tasks. On the system side, we investigate application-level strategies to effectively mitigate domain shift in real-world settings, including a Latent Replay mechanism to prevent catastrophic forgetting in a Continual Learning setup for image classification and a multi-modal label generation scheme for autonomous ODL in monocular depth estimation. Through this system-level approach, we demonstrate that effective on-device adaptation can be achieved in under 20 minutes for both application domains. Specifically, a MobileNetV2 model for image classification learns a new class in just 18 minutes—1.3× faster than a full-model update—using only 4.63 MB of memory for the Continual Learning task, while a lightweight μPyD-Net model for monocular depth estimation can adapt to an unseen domain in 17.8 minutes using 3 k self-acquired and self-labeled images and less than 1.2 MB of memory to update its weights. These results establish the practical viability of ODL across diverse applications, paving the way for a new generation of self-adapting, autonomous, and pervasive ULP IoT devices.