**On-Device Learning (ODL) on RISC-V Multicore MCUs**

Davide Nadalini, Manuele Rusci, Luca Benini, and Francesco Conti

1 Politecnico di Torino, Torino, Italy, 2 Università di Bologna, Bologna, Italy, 3 Katholieke Universiteit Leuven, Leuven, Belgium, 4 ETH Zurich, Zurich, Switzerland

**Motivation**

Many DNNs rely on Convolutions trained with Floating-Point BackPropagation. On-Device Inference & Adaptation limited by MCU performances / resources.

**Our Goal:** enabling ODL on ultra-low-power MCUs

**Our target:** the PULP Platform!

1. **PULP-TrainLib:** Enabling On-Device Training for RISC-V Multicore MCUs through Performance-Driven Autotuning

2. **PULP-TrainLib:** the first open-source training library for RISC-V multicore MCUs with Matrix-Multiplication (MM)-based performance-tunable Floating-Point (FP) primitives.

**BackPropagation-based ODL primitives of CNN models**

**DNN Layer**

- **PULP-TrainLib:** convolutional platform for energy-efficient and scalable edge computing based on RISC-V cores.

- **Table 1:** PULP-TrainLib's optimized MM algorithms

**Workhorse:** FP32 Unrolled & Parallel MM Library

**Example:** CHW Conv2D

Matrix Multiplication based ODL steps:

<table>
<thead>
<tr>
<th>Step</th>
<th>PULP Settings</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Layer type size &amp; step</td>
<td>PULP SoC Setup (L, K, N) Conv2D</td>
<td>Layer is tiled if exceeds L1, exhaustive search with HIL (PULP GSoC simulation) for optimal setup</td>
</tr>
<tr>
<td>Layer type size &amp; step</td>
<td>PULP GSoC Setup: L1 size = 64 kB, Ncores = (1, 8)</td>
<td></td>
</tr>
</tbody>
</table>

**Layer type size & step:**

- Up to 2x speedup (autotuned vs one-size-fits-all, 4.39 MAC/clock on 8 RISC-V cores)
- 36.6x less latency vs unoptimized STM32

**PULP GSoC Setup:**

L1 size = 64 kB, Ncores = (1, 8)

**Optimal tile size & MM algorithm for layer setup**

**DNN Layer & Step**

- Layer is tiled if exceeds L1, exhaustive search with HIL (PULP GSoC simulation) for optimal setup

**Layer type size & step**

- PULP GSoC Setup: L1 size = 64 kB, Ncores = (1, 8)

**3:** a detailed analysis of PULP-TrainLib and AutoTuner on an 8-Core PULP Platform

- PULP GSoC Setup: L1 size = 64 kB, Ncores = (1, 8)

- Optimized HWC Primitives

- E.g., FP16 Conv2D:
  - Weights stored in transposed form
  - Primitive matrix expression reshaped to exploit MM kernels

**Ideal estimation of a single training step of two TinyML Perf models**

- With (tilling and AutoTuner)

**FC6:**

- 34.51% faster than CHW

- FP16 1.9x faster than FP32

**Conclusion**

- Enabling ODL on MCUs allows real-time BackProp-based learning on IoT end nodes (Continual, Online, ... Learning)
- Reduced Precision enables fast (1.86x FP32) backend for ODL with high enough precision
- Power consumption < 100 mW on Multi-Core RISC-V MCUs

- Latent Replay for Real-Time Continual Learning: [webpage]