WIP: Automatic DNN Deployment on Heterogeneous Platforms: the GAP9 Case Study
Energy-Efficient Embedded Systems Laboratory (UNIBO)

Luka Macan <luka.macan@unibo.it>
Alessio Burrello <alessio.burrello@polito.it>
Luca Benini <luca.benini@unibo.it>
Francesco Conti <f.conti@unibo.it>

PULP Platform
Open Source Hardware, the way it should be!
Goal: NN Deployment on the Edge

Our target: NN accelerated embedded edge devices

• Heterogenous
• Ultra-low-power
• Software managed data "caches"

Case study: GAP9 SoC

• Compute: Cluster of 8 RISC-V cores + NN Accelerator
• 0.33 mW/GOP\(^1\)
• 3 levels of memory + DMAs

\(^1\)https://greenwaves-technologies.com/gap9_processor/
**Inter-layer optimizations**

- **Software pipelining & Task parallelization**
  - Core 1: Load 0, Prep job 0, Load 1, Prep job 1, Load 2
  - Core 2: Acc. call 0
  - Core 3: Acc. call 1, Store 0
  - Core 4: Job 0, Job 1

- **Tiling heuristics**
  1. Tiles divisible with input buffer
  2. Balanced tiles
  3. Full dimensions
Inter-layer optimizations

- Software pipelining & Task parallelization
- Tiling heuris

12% overall latency overhead over ideal
Intra-layer optimizations

- Layer fusion

- Choice of layers to fuse
Intra-layer optimizations

- Layer fusion
- Choice of layers to fuse

MN-V2 unique Inverted Bottleneck blocks
Conclusion

• Achieved high accelerator utilization (~90%) and 3.44x performance improvement over the predecessor, GAP8

• Accelerators need several techniques to be applied for them to come close to ideal utilization

• Automatic deployment tools will play a more and more critical role as the hardware gets more complex