Aggressive Data Transfer Reduction by Loop Fusion and Re-Computation for Dedicated Accelerators

Maurice Peemen, Bart Mesman, and Henk Corporaal

So-called "Smart Companion Devices" are Not Intelligent Yet!

Many smart companion devices, but all without intelligence. Deep Neural Nets and Machine Learning brings Superhuman recognition accuracy [1, 2] on complex tasks to wearables.

Huge data transfer requirements prevent efficient implementation of such intelligent functionalities with acceptable battery lifetime.

HLS Based Accelerators for Energy Efficiency

Demanding apps require highly customized accelerators to achieve challenging throughput and energy constraints. Energy is often dominated by data movement and operator energy is neglectable. State-of-the-art uses loop tiling and interchange [3, 4] to optimize data reuse in local accelerator buffers.

Looking Beyond Tiling: Fusion and Re-Computation

Tiling or classical fusion with row buffers does not work well, limited reuse or large buffer requirements. Introducing redundant workload could help out.

Design Space Exploration

Manual exploration intractable, therefore automate the search through solution space. We obtain optimized schedules regarding locality and buffer size.

The Fruits of Our Approach

Apply the suggested transformations to the filters and synthesize with HSL, mapping results to Virtex 6 FPGA at 150 MHz.

- Fast Design Space Exploration (seconds)
- Up to 2x communication reduction w.r.t. [4]
- Up to 8x better throughput

References