WIP: AUTOMATIC DNN DEPLOYMENT ON HETEROGENEOUS PLATFORMS: THE GAP9 CASE STUDY

Luka Macan*, Alessio Burrello, Luca Benini, Francesco Conti
*luka.macan@unibo.it

Background

Target systems:
- Heterogeneous
- Embedded
- Low-power

Problem

Data movement overhead

Control overhead

Layer size preferences

Applied solution

Software Pipelining

Task-Level Parallelism

Tiler Heuristics

Result 1: With above mentioned techniques, achieved 12% overall latency overhead over ideal execution on MobileNetV1

Result 2: Layer Fusion is not a one-shoe-fits-all solution, and the choice of layers is important

Result 3: Achieved 91% and 89% accelerator utilization on MobileNet-V1 and MobileNet-V2 respectively. Speedup of 3.44x over DORY on GAP8 (no acc.).

NE16 overview

DORY
- DNN deployment
- Parallel ultra-low-power platforms
- Tiling ILP
- Open-source

Inter-Layer

Layer Fusion

Data Movement Overhead

Layer 1

Layer 2

<table>
<thead>
<tr>
<th>Perf. (MAC/cycle)</th>
<th>Peak-Perf. Perc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAP8 MN-V1</td>
<td>8.09</td>
</tr>
<tr>
<td>GAP9 MN-V1</td>
<td>26.04</td>
</tr>
<tr>
<td>GAP9 MN-V2</td>
<td>24.23</td>
</tr>
</tbody>
</table>

Acknowledgements: Funded by ROADSTER and NeuroSoC (g.a. 101070634) projects