Benchmarking AI-methods on Heterogeneous Hardware Resources

Christopher Hesse, Holger Eichelberger

{eichelberger}@sse.uni-hildesheim.de,
hessech@uni-hildesheim.de

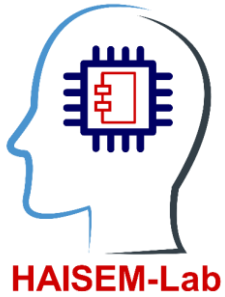
Software Systems Engineering
University of Hildesheim

www.sse.uni-hildesheim.de

HAISEM-Lab

Motivation

- Artificial Intelligence (AI) is “everywhere”
- HAISEM-Lab (http://haisem-lab.de/)
  - BMBF founded AI lab
  - **H**ardware-optimized **A**rtificial **I**ntelligence Applications using modern **S**oftware **E**ngineering **M**ethods
  - Qualification and training for industry personnel
  - AI/Hardware/SE research
- Partners
  - University of Hannover (L3S, IMS)
  - University of Hildesheim (SSE)



Problem

- Hardware for AI
  - GPU server 8 NVIDIA Tesla

  - FPGA server with 2 Maxeler Maia cards

  - TPU/GPU developer boards, GPU laptops

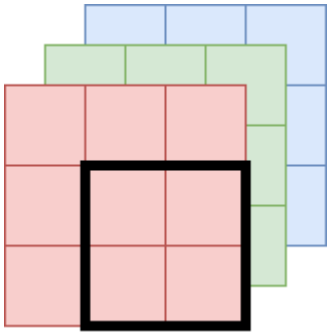
*How to compare AI performance (speed, energy) across **all** heterogeneous hardware resources?*

- Existing approaches: At least one hardware type missing

Approach

- Focus: Convolutional Neural Networks
  - "What's in this image?“

- Bottom-Up
  - Microbenchmarks
    - Convolution
    - Pooling
  - Macrobenchmarks
    - Training
    - Inference



- Methodology
  - Varying input/filter size, e.g., 100x100, 1000x1000, …
  - Run each benchmark for *n* seconds
  - Measure per iteration / benchmark
    - CPU/GPU time
    - Energy

Preliminary results

2D Convolution	Laptop (CPU)	Laptop (GPU)	Server (CPU)	Server (GPU)
Speedup	1.0x	**~ 10-20x**	1.0x	**~ 10-75x**
GPU Power			45 W	65-295 W
Active Power			760-850 W	760-1150 W

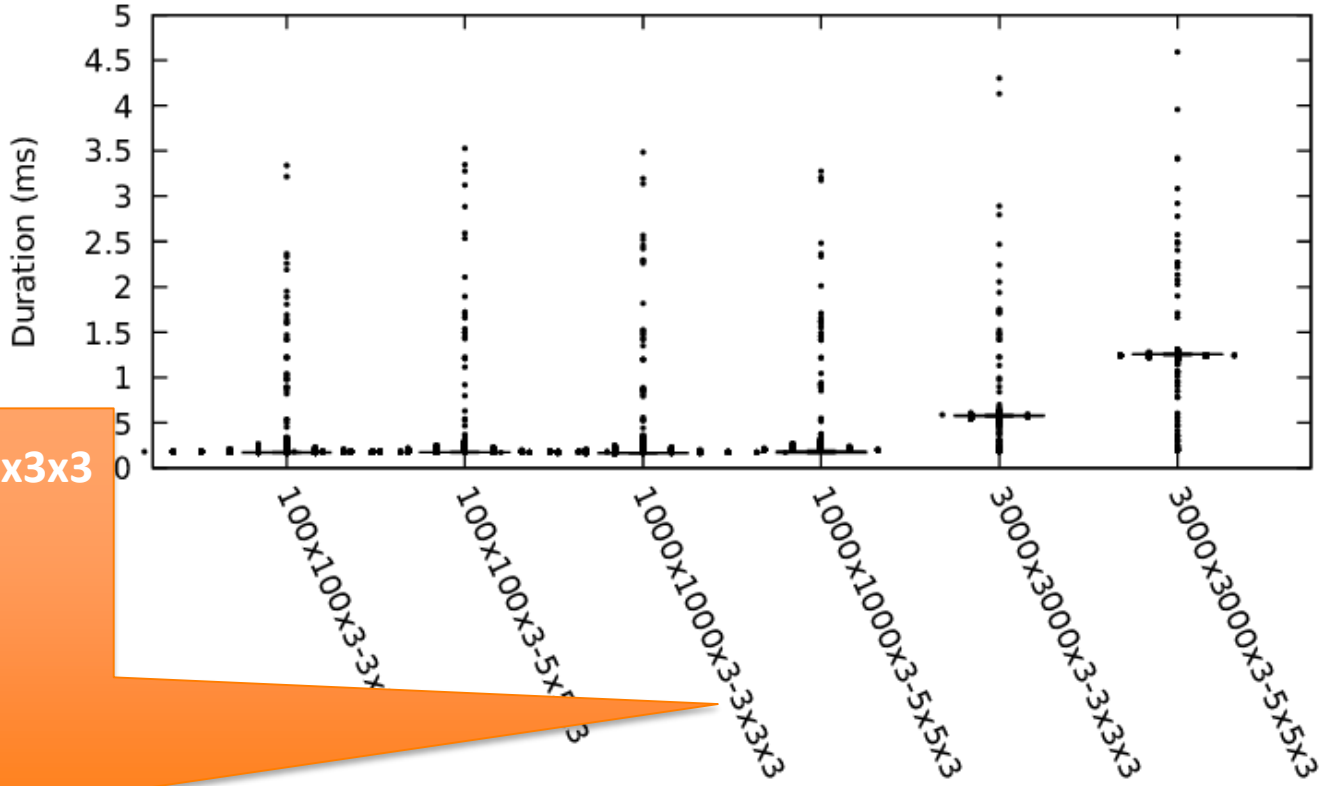
CNN Training 32x32	Laptop (CPU)	Laptop (GPU)	Server (CPU)	Server (GPU)
Speedup	1.0x	~ 1.2x	1.0x	~ 1.6x
Memory	~ 4.5 GB	~ 5.6 GB	~ 5.0 GB	~ 8.3 GB

Input size too small? 224x224 scales much better.

Preliminary results

2D Convolution (DGX-1 Server, GPU)

- Width x Height x Channels



Discussion: 1000x1000-3x3x3

t = 60 seconds
n = 320.000
σ = 1,8 x 10^{-5}

Outliers: ~2,4% (random)

Conclusions & Future Work

- Microbenchmarks: good scaling with hardware capability
- Macrobenchmarks: it depends ;)

- Compare with more/less GPUs
- Realize micro-benchmarks on FPGA
- Compare with existing benchmarks where possible
- Derive “best practice” tradeoffs

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

HAISEM-Lab
http://haisem-lab.de/