RECHNER-
ARCHITEKTUR
INFIII

FAU
FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
FACULTY OF ENGINEERING

RADL – Rechnerarchitekturen für Deep-Learning

Philipp Holzinger, Philipp Gündisch

Introduction

Organization

Topic Suggestions

Introduction

Organization

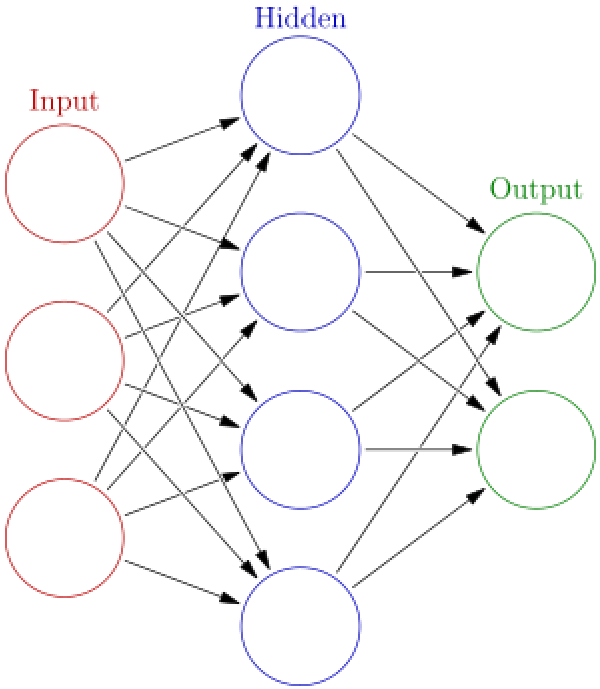
Topic Suggestions

Deep Learning

- **Part of AI**
- **Good if:**
  - Complex rules with hand optimized algorithms needed
  - Complex problems without traditional approach
  - Fluctuating environment
  - To learn more about the problem itself
- **Lots of research ongoing**
  - Algorithms, Nets, Learning
  - HW Architectures

Deep Learning

- **Applications:**
  - Image processing
  - Speech recognition
  - „All tasks, that are easy for humans, but difficult for algorithms“
- **Basic Principle**
  - Deep Neuronal Networks
  - CNN, RNN, LSTM, …
- **Here:**
  - Focus on inference (not training)
  - Hardware architectures for this task



Hardware Architectures

- **What is the best architecture for learning/inference of artificial neural networks (ANN)?**
  - CPU
  - GPU
  - FPGA
  - Dedicated ASIC

- **How to use architectural properties?**
  - SIMD, OpenMP, Multithreading
  - OpenCL, Partitioning, Warps, SIMT
  - HLS, Parallelism, Interfacing
  - Algorithmic Optimization for different HW architectures

Introduction

Organization

Topic Suggestions

Organization

- **Tutors: Philipp Holzinger, Philipp Gündisch**

- **Reminder: This is not a lecture but a practical course!**
  - No regular lectures where we teach you about topics (only selected ones)
  - Instead: everybody works on own project and we guide you through it
  - Everybody researches the required knowledge about the project by themselves
  - Everybody informs the other participants about their findings in regular sessions

Organization

- **Results:**
  - Final presentation (50%)
  - Architecture report (50%)
- **Bachelor / Master: 10 ECTS**
- **Join StudOn Course!**
  - **Name: Rechnerarchitekturen für Deep-Learning Anwendungen**

Organization

- **Procedure**
  1. Intro – *today*
  2. DL fundamentals – *today*
  3. Introduction into architectures – *next week*
  4. Presentation of project proposals – *in 2 weeks*
  5. Project in groups – *until end of semester*
     1. Small groups (2-3 students) investigate ONE architecture
     2. Every second week discussion round with small talks about your findings
     3. Individual meetings with tutors if needed
  6. Consolidation – *end of semester*
     1. Each group gives a final talk about their architecture
     2. Each group writes a report about their architecture

Organization

- **Projects about are the realization, optimization and evaluation of deep learning algorithms on different architectures (choose ONE)**
  1. CPU x86/ARM (server / embedded)
  2. GPU (server / embedded)
  3. Dedicated circuit on FPGA (VHDL / HLS C)
  4. Coral Edge TPU

- **If needed you can loan an embedded device for home-use**
  - Please make sure you have an insurance

Organization

- **What to do now?**
  - Make yourself familiar with one of the common DL Frameworks
    - Train a network that you can use for the remainder of the course
    - learn how to export its weights into an easily readable format
  - Find team partners and think about the algorithms and architecture you want to research

- **Next Meetings**
  - 16.10.2024 – introduction to deep learning
  - 23.10.2024 – introduction to architectures
  - 30.10.2024 – presentation of project proposals
  - 06.11.2024 – no meeting

Introduction

Organization

Topic Suggestions

Topic Suggestions

CPU: matrix multiplication and convolution with threads, SIMD, cache effects – problem size & optimization

GPU: workgroup-sizes, shared memory, PCIe latency

FPGA: (sparse) matrix multiplication and convolution accelerator (with VHSL or HLS), resources vs. throughput vs. latency

TPU: usage, performance, efficiency