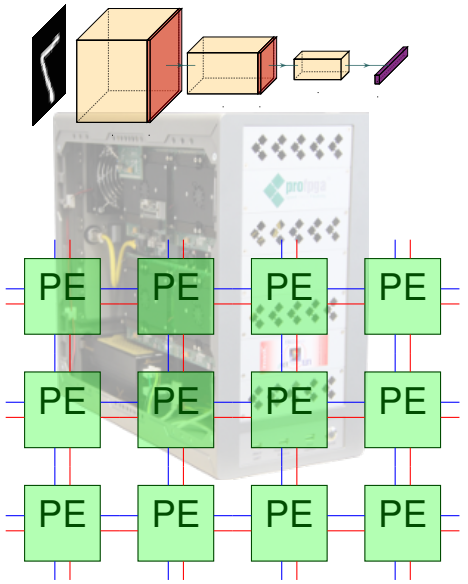
Mapping of CNNs onto FPGA Overlays

Deep Neural Networks are increasingly employed in real-time embedded systems, often using a Programmable System-on-Chip (PSoC) including a Field Programmable Gate Array (FPGA), which can be utilized for hardware acceleration. For easy deployment on FPGAs, often hardware overlays for common operations of the applications are used.
In this work, a PSoC is used for accelerating a Convolutional Neural Network (CNN) for object detection to control the grasping force of a prosthetic hand. For the CNN acceleration, *Tightly Coupled Processor Arrays* (TCPAs), a hardware blueprint, developed at our chair, consisting of a reconfigurable 2D grid of processors is deployed on an FPGA enabling fast acceleration of the compute-intensive layers of a CNN (i.e., convolutional layers). The work consists of the following tasks:



- Map CNN layers to a TCPA architecture on an FPGA prototyping platform.
- Comparison with another popular hardware overlay, the Versatile Tensor Accelerator (VTA), regarding area efficiency and execution time for different layers for an exemplary CNN.

Prerequisites:	Basic knowledge about CNNs, Python, and VHDL; Programming skills in C++
Type of work:	Theory (25%), conception (25%), implementation (50%)
Supervisor:	Christian Heidorn (christian.heidorn@fau.de)

Lehrstuhl für Informatik 12
Hardware-Software-Co-Design
Cauerstraße 11
91058 Erlangen

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT