Acceleration of Deep Learning for Cloud and Edge Computing

Project status: current
In this project, we explore efficient algorithms and architectures for state-of-the-art deep learning based applications. The first work, Caffeine, offers a uniformed framework to accelerate the full stack of convolutional neural networks (CNN), including both convolutional layers and fully-connected layers. Following the first work, we further explore the efficient microarchitecture for implementing the computation-intensive kernels in CNN. A special architecture, systolic array, which consists of processing elements (PEs) with local interconnects, are thoroughly studied. Meanwhile, in the project of CLINK, a LSTM inference kernel is designed for EEG signal processing on neurofeeedback devices, which demonstrates high speedups and energy efficiency on FPGAs compared to CPU and GPU counterparts.

Below are the detailed summaries of each project.

1. Caffeine: Towards Uniformed Representation and Acceleration for Deep Convolutional Neural Networks

With the recent advancement of multilayer convolutional neural networks (CNN), deep learning has achieved amazing success in many areas, especially in visual content understanding and classification. To improve the performance and energy-efficiency of the computation-demanding CNN, the FPGA-based acceleration emerges as one of the most attractive alternatives. We design and implement Caffeine, a hardware/software co-designed library to efficiently accelerate the entire CNN on FPGAs. First, we propose a uniformed convolutional matrix multiplication representation for both computation-intensive convolutional layers and communication-intensive fully connected (FCN) layers. Second, we design Caffeine with the goal to maximize the underlying FPGA computing and bandwidth resource utilization, with a key focus on the bandwidth optimization by the memory access reorganization not studied in prior work. Moreover, we implement Caffeine in the portable high-level synthesis and provide various hardware/software definable parameters for user configurations. Finally, we also integrate Caffeine into the industry-standard software deep learning framework Caffe.

2. Automatic Systolic Array Synthesis

Modern FPGAs are equipped with an enormous amount of resource. However, existing implementations have difficulty to fully leverage the computation power of the latest FPGAs. We implement CNN on an FPGA using a systolic array architecture, which can achieve high clock frequency under high resource utilization. We provide an analytical model for performance and resource utilization and develop an automatic design space exploration framework, as well as source-to-source code transformation from a C program to a CNN implementation using systolic array. The experimental results show that our framework is able to generate the accelerator for real-life CNN models, achieving up to 461 GFlops for floating point data type and 1.2 Tops for 8-16 bit fixed point.

The project above works on the systolic array synthesis for CNN. We are also working on improving the generability of the approach to map more general applications to systolic arrays. We present our ongoing compilation framework named PolySA which leverages the power of the polyhedral model to achieve the end-to-end compilation for systolic array architecture on FPGAs. PolySA is the first fully automated compilation framework for generating high-performance systolic array architectures on the FPGA leveraging recent advances in high-level synthesis. We demonstrate PolySA on two key applications—matrix multiplication and convolutional neural network. PolySA is able to generate optimal designs within one hour with performance comparable to state-of-the-art manual designs.

3. CLINK: Compact LSTM Inference Kernel for Energy Efficient Neurofeedback Devices

Neurofeedback device measures brain wave and generates feedback signal in real time and can be employed as treatments for various neurological diseases. Such devices require high energy efficiency because they need to be worn or surgically implanted into patients and support long battery life time. In this paper, we propose CLINK, a compact LSTM inference kernel, to achieve high energy efficient EEG signal processing for neurofeedback devices. The LSTM kernel can approximate conventional filtering functions while saving 84% computational operations. Based on this method, we propose energy efficient customizable circuits for realizing CLINK function. We demonstrated a 128-channel EEG processing engine on Zynq-7030 with 0.8 W, and the scaled up 2048-channel evaluation on VirtexVU9P shows that our design can achieve 215x and 7.9x energy efficiency compared to highly optimized implementations on E5- 2620 CPU and K80 GPU, respectively. We carried out the CLINK design in a 15-nm technology, and synthesis results show that it can achieve 272.8 pJ/inference energy efficiency, which further outperforms our design on the Virtex-VU9P by 99x.

Publications:

Caffeine: Towards Uniformed Representation and Acceleration for Deep Convolutional Neural Networks (/publications/caffeine-towards-uniformed-representation-and-acceleration-deep-convolutional-neural)

Optimizing FPGA-based Accelerator Design for Deep Convolutional Neural Networks (/publications/optimizing-fpga-based-accelerator-design-deep-convolutional-neural-networks)

Automated Systolic Array Architecture Synthesis for High Throughput CNN Inference on FPGAs (/publications/automated-systolic-array-architecture-synthesis-high-throughput-cnn-inference-fpgas)

PolySA: Polyhedral-Based Systolic Array Auto-Compilation (/publications/polysa-polyhedral-based-systolic-array-auto-compilation)

CLINK: Compact LSTM Inference Kernel for Energy Efficient Neurofeedback Devices (/publications/clink-compact-lstm-inference-kernel-energy-efficient-neurofeedback-devices)

Faculty:

Jason Cong (/people/faculty/jason-cong)

Students:

Chen Zhang (/people/visiting-researcher/chen-zhang)
Peipei Zhou (/people/student/peipei-zhou)
Jie Wang (/people/student/jie-wang)
Hao Yu (/people/student/hao-yu)