Oral S03

Digital Circuit Design and High-Level Sythesis for AI Acceleration

Date/Time	8/2(三)　13:30-15:00
Chair(s)	黃俊達／國立陽明交通大學電機工程學系 施信毓／國立中山大學電機工程學系

S03.1 | 13:30—13:45

Design and Implementation of Hardware-Friendly Activation Functions for Transformer Applications

Yu-Hsiang Huang, Pei-Hsuan Kuo, Ting-Wei Hu, and Juinn-Dar Huang
Department of Electronics and Electrical Engineering & Institute of Electronics
National Yang Ming Chiao Tung University

The activation function is one of key elements in machine learning algorithms. However, some broadly-used activation functions are exceptionally complex, e.g., GELU in Transformer-based algorithms, which makes their precise yet efficient VLSI implementations extremely hard. In this paper, two series of hardware-friendly activation function designs, DNR and PWL, and their VLSI implementations are proposed. Both are specifically designed to replace GELU, which is widely used in Transformer-related applications. Instead of utilizing traditional lookup-table (LUT)-based approximation methods, this paper introduces new activation functions that are not only hardware-friendly but successfully alleviate the dying neuron issue. Besides, each series includes a number of members, which can be freely selected through programming to best fit a given application. Experimental results indicate that the proposed new activation functions achieve comparable or even better model accuracy as compared to GELU. Moreover, the highly efficient and flexible VLSI implementations support 16 different Q-formats to maximize the output precision under various input scales. Compared with approximation-based implementation strategies, the proposed activation function designs and the corresponding LUT-free hardware implementations do achieve a significant improvement in speed, area, and power.

S03.2 | 13:45—14:00

Accelerating CNNs for Particle Energy Reconstruction on FPGAs

Chi-Jui Chen [1], Yan-Lun Huang [2], Ling-Chi Yang [3], Ziang Yin [4], Bo-Cheng Lai [3], Scott Hauck [4], Shih-Chieh Hsu [5], Philip Harris [6], Dylan Rankin [7]

[1] Graduate Degree Program of College of Electrical and Computer Engineering, National Yang Ming Chiao Tung University
[2] Department of Electrical and Computer Engineering, National Yang Ming Chiao Tung University
[3] Institute of Electronics, National Yang Ming Chiao Tung University
[4] Department of Electrical and Computer Engineering, University of Washington
[5] Department of Physics, University of Washington
[6] Laboratory for Nuclear Science, Massachusetts Institute of Technology
[7] Department of Physics and Astronomy, University of Pennsylvania

In recent years, the Large Hadron Collider (LHC) experiments at CERN have incorporated deep learning (DL) techniques to enhance the quality of data analysis. However, the tremendous data rate and energy consumption result in a severe gap between local processing and real-time inference. In this paper, we present the first fully-automated design and optimization workflow based on hls4ml to deploy large convolutional neural networks (CNNs) on FPGAs. Considering DeepCalo models trained on ATLAS data, we show that optimized stream-based dataflow can efficiently benefit from the fully-on-chip deployment. The 1.8-millions-parameters model inferences at a latency of 0.443 ms, 14.1x faster than using a Tesla V100 GPU. We study the impacts of applying different quantization schemes to fit the computational and latency constraints. We further explore several rounding strategies used in hardware to eliminate the quantization error during the early stage. At last, we compare the FPGA implementation with other co-processors using various indicators.

S03.3 | 14:00—14:15

An Efficient Sparse CNN Architecture with Index-based Kernel Transformation

Fang-Yi Gu, Po-Ting Chen, Shan-Chi Yu, and Ing-Chao Lin
Department of Computer Science and Information Engineering, National Cheng Kung University

In order to reduce the memory requirements and energy consumption of convolutional neural networks (CNNs), this paper proposes an index-based kernel transformation implemented on networks whose weights are quantized into an index-based representation while keeping the fixed-point precision. The proposed algorithm can eliminate redundant operations by extracting the common index patterns from different kernels to perform the identical operation only once. A specifically designed hardware is implemented to perform the convolution and rebuild the correct results from

extracted patterns in parallel. The experiment shows that deploying a large network like VGG-16 on the proposed hardware only requires 2.620KB of on-chip memory and has energy efficiency better than the state-of-the-art.

S03.4 | 14:15—14:30

A Lossless Winograd Parallel Convolutional Neural Networks Accelerator

Hu-Chun-Yi Hsieh, Chung-Bin Wu
National Chung Hsing University

To accelerate the computation speed of CNNs to meet the demands of real-time processing, the need for accelerator designs emerged. This paper designed a CNN accelerator with configurable parallelism to adapt to different networks and environments. The proposed accelerator incorporates the Window Combine method to improve the efficiency of edge computations with zero-padding. It achieves a performance improvement of 9.44% on VGG16.The implementation of VGG-16 on the ZCU102 FPGA platform demonstrates that the accelerator achieves a good throughput of 228.6 GOP/s. The DSP Efficiency is 0.89 GOP/s/DSPs, and the Memory Efficiency is 2.01GOP/s/KB.

S03.5 | 14:30—14:45

Configurable Multi-Precision Floating-Point Multiplier Design Optimized for Inferences in Deep Learning Applications

Pei-Hsuan Kuo, Yu-Hsiang Huang, Jye-En Wu, and Juinn-Dar Huang
Department of Electronics and Electrical Engineering & Institute of Electronics, National Yang Ming Chiao Tung University

The increasing AI applications demands efficient computing capabilities to support a huge amount of calculations. Among the related arithmetic operations, multiplication is an indispensable part in most of deep learning applications. To support computing in different precisions demanded by various applications, it is essential for a multiplier architecture to meet the multi-precision demand while still achieving high utilization of the multiplication array and power efficiency. In this paper, a configurable multi-precision FP multiplier architecture with minimized redundant bits is presented. It can execute 16× FP8 operations, or 8× brain-floating-point (BF16) operations, or 4× half-precision (FP16) operations, or 1× single-precision (FP32) operation every cycle while maintaining a 100% multiplication hardware utilization ratio. Moreover, the computing results can also be represented in higher precision formats for succeeding high-precision computations. The proposed design has been implemented using the TSMC

40nm process with 1GHz clock frequency and consumes only 16.78mW on average. Compared to existing multi-precision FP multiplier architectures, the proposed design achieves the highest hardware utilization ratio with only 4.9K logic gates in the multiplication array. It also achieves high energy efficiencies of 1212.1, 509.6, 207.3, and 42.6 GFLOPS/W at FP8, BF16, FP16 and FP32 modes, respectively.

S03.6 | 14:45—15:00

Design and Exploration of A Cluster-Based DNN Accelerator with High-Level Synthesis

Jyun-Siou Huang[1] , Ting-Han Chou[2] , Juin-Ming Lu[2] , Chih-Tsun Huang[1] , and Jing-Jia Liou[2]
[1] Department of Computer Science, National Tsing Hua University
[2] Department of Electrical Engineering, National Tsing Hua University

With the increasing intensity of deep neural network (DNN) workloads, developing various hardware accelerators and their estimation tools have become vital to tackle this challenge. Consequently, the design of interconnections has emerged as a crucial component in achieving high performance when utilizing clusters of processing elements in the accelerator. This paper proposes a novel architecture of cluster-based accelerator architecture with high-level synthesis, enabling rapid exploration of the design space for scalable interconnection bandwidth and cluster size. Additionally, we leverage tag data and work modes to introduce flexibility in data communication without restricting the type of dataflow. Through the evaluation, our proposed architecture demonstrates performance improvements ranging from 14\% to 67.8\% for different neural network models, achieved simply by adjusting the cluster size.