論 文 内 容 の 要 旨

博士論文題目

A Fully-Pipelined Inference Accelerator for Deep Convolutional Neural Networks
(深層畳み込みニューラルネットワーク向け完全パイプライン化推論アクセラレータ)

氏 名 NGUYEN VAN CAM

Due to the high speed and power efficiency of the field-programmable gate array (FPGA), many FPGA-based inference accelerators for deep convolutional neural network (CNN) have been widely adopted. Large-scale CNNs require intensive computations as well as a large amount of storage space and memory access. However, low bandwidth off-chip memories are a main challenge for data transmission between external memory and FPGA-based CNN inference accelerator.

In this research, we develop the following to improve performance and power efficiency. First, we use a high bandwidth memory (HBM) to expand the bandwidth of data transmission between the off-chip memory and the accelerator. Second, a fully-pipelined manner, which consists of pipelined inter-layer computation and a pipelined computation engine, is implemented to decrease idle time among layers. Third, a multi-core architecture with shared-dual buffers is designed to reduce off-chip memory access and maximize the throughput.

We designed the proposed accelerator on the Xilinx Alveo U280 platform with in-depth Verilog HDL instead of high-level synthesis as in the previous works and explored the VGG-16 model to verify the system during our experiment. With a similar accelerator architecture, the experimental results demonstrate that the memory bandwidth of HBM is 13.2× better than DDR4. Compared with other accelerators in terms of throughput, our accelerator is 1.9×/1.(5×/11.9× better than FPGA+HBM2 based / low batch size GPGPU / low batch size CPU. Compared with the previous DDR+FPGA / DDR+GPGPU / DDR+CPU based accelerators in terms of power efficiency, our proposed system provides 1.4-1.7×/1.7-12.6×/6.6-37.1× improvement with the large-scale CNN model.

氏　名	NGUYEN VAN CAM

（論文審査結果の要旨）（Ａ４　１枚　１、２００字程度）

FPGA の高速性と電力効率を利用する、深層畳み込みニューラルネットワーク(CNN)向け FPGA ベース推論アクセラレータが、数多く研究開発されている。しかし、大規模 CNN には、大量のストレージおよびメモリに加えて、膨大な計算が必要である。そして、低帯域幅のオフチップメモリが、FPGA ベース推論アクセラレータにおける、深刻な性能ボトルネックとなっている。

以上の課題に対して、本研究は、性能と電力効率を向上させるために、以下を提案している。第１に、高帯域幅メモリ(HBM2)を使用して、オフチップメモリとアクセラレータ間のデータ転送帯域幅を拡張した。 第２に、パイプライン化されたレイヤ間計算と、パイプライン化された計算エンジンにより構成される、完全パイプライン化システムを実装し、レイヤ間のアイドル時間を短縮している。第３に、共有デュアルバッファを備えたマルチコアアーキテクチャにより、オフチップメモリアクセスを削減し、スループットを最大化している。また、先行研究に多く見られる、高級言語による記述と高レベル合成ではなく、ハードウェアの詳細記述が可能な Verilog HDL を使用して、Xilinx Alveo U280 プラットフォーム上で稼働する実システムを開発し、また、小規模 CNN だけでなく、VGG-16 による実用的モデルも評価している。さらに、メモリのみを変更したアクセラレータとの比較により、HBM2 の基本メモリ帯域幅が DDR4 の 13.2 倍優れていることを示している。先行研究のアクセラレータと比較した場合、本アクセラレータは、スループットの観点から、FPGA+HBM2 ベース / 低バッチサイズ GPGPU / 低バッチサイズ CPU よりも 1.9 倍/1.65 倍/11.9 倍優れている。電力効率の観点から、大規模 CNN モデルでは、従来の DDR+FPGA / DDR+GPGPU / DDR+CPU ベースアクセラレータと比較して、1.4-1.7 倍 / 1.7-12.6 倍 / 6.6-37.1 倍改善している。

以上、本論文は学術上、実際上寄与するところが少なくない。よって、本論文は博士（工学）の学位論文として価値あるものと認める。