論文 / 著書情報
Article / Book Information

題目(和文)	FPGAベースの機械学習アクセラレータの設計最適化に関する研究
Title(English)	A Study on Design Optimization for FPGA-based Machine Learning Accelerator
著者(和文)	神宮司明良
Author(English)	Akira Jinguji
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第11761号, 授与年月日:2022年3月26日, 学位の種別:課程博士, 審査員:中原 啓貴,髙橋 篤司,本村 真人,劉 載勲,佐々木 広,高前田 伸也
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第11761号, Conferred date:2022/3/26, Degree Type:Course doctor, Examiner:,,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	情報通信 系 情報通信 コース	申請学位（専攻分野）： Academic Degree Requested	博士 （ 工学 ） Doctor of
学生氏名： Student's Name	神宮司明良	指導教員（主）： Academic Supervisor(main)	中原啓貴 准教授
		指導教員（副）： Academic Supervisor(sub)	

要旨（英文 800 語程度）
Thesis Summary（approx.800 English Words ）

An autonomous system has several requirements: high recognition accuracy, real-time response time, low power consumption, and low manufacturing cost. There are several computing methods to achieve this, especially in embedded devices, the most typical of which is the CPU. CPUs in embedded devices are inexpensive and have low power consumption. However, their performance may be insufficient for real-time response. Therefore, the use of processors explicitly designed for the application, such as GPUs, ASICs, and FPGAs, can be considered. GPUs have high computational performance for matrix computation, but their power consumption is high, and they are not suitable for flexible computation, and designing dedicated circuits with ASICs requires a very high cost for design. We are interested in FPGAs because they are capable of very flexible computation and have a good balance between moderately high-performance and relatively low design costs. FPGAs have a very high potential due to the miniaturization of semiconductor process size and hard macros. We also believe that by studying the excellent design of FPGAs, it is possible to exploit their high potential in terms of computing performance, power consumption, and design cost. This research proposes an optimization design that satisfies all the constraints of computational performance, power consumption, and design cost on machine learning in autonomous systems. However, the following problem still exist; as for processing speed, an FPGA allows small and flexible circuit design, and its processing speed can be improved by co-designing with the machine learning algorithm itself; as for power consumption, there exists a trade-off between performance and power consumption; as for the design cost, it lacks a design flow for a low-cost device for machine learning. We are paying attention to FPGAs with high potential to meet all constraints with the above improvements. Our optimization method will bring us one step closer to realizing autonomous systems and accelerate the arrival of newcomers. The objective of this research is to design an FPGA-based machine learning accelerator that satisfies the three constraints of inference speed, power consumption, and design cost. At first, we will focus on random forests and propose to design a random forest using HLS. In this chapter, to further reduce the amount of hardware, we propose an optimization that uses k-means clustering to share the comparators of branching nodes on the decision tree. We have improved the trade-off between recognition accuracy and hardware usage for random forests by sharing the thresholds. We have reduced the design cost by proposing a series of design flows from model training to threshold sharing and hardware design for random forests. We implemented this random forest on a Xilinx FPGA and achieved a speedup of more than 8.4 times compared to the conventional method. Secondly, we will focus on more practical tasks such as human pose estimation and CNNs, which have higher recognition accuracy. Since CNNs are computationally intensive, parallelization is not enough to speed up inference in real-time. To achieve faster inference, we focused on the sparse weight CNNs. We have achieved about 3.5 times faster inference speed and about 13 times better power efficiency compared to the existing GPU method. Thirdly, we discuss how to realize CNNs on FPGA devices with smaller power consumption. We use Split-CNN for solving feature-map size problem. We have clarified the trade-off between recognition accuracy and hardware usage by splitting the Split-CNN. We designed a memory buffering method and scheduling for Split-CNN and implemented it on a PYNQ-Z1 FPGA board, a low-end FPGA with a power consumption of about 3W. With this achievement, we have succeeded in reducing the power consumption to less than one-third while achieving 3.1 times faster speed compared to GPU. Finally, we will consider using multiple small circuit modules designed using high-level synthesis to achieve a high operating frequency. We propose an architecture using a ring bus and computation scheduling for CNNs to hide the communication overhead caused by partitioning the circuits. When implemented on a Xilinx FPGA board ZCU102, we were able to achieve an operating frequency of 500MHz despite the relatively large design occupying more than 140k LUTs and 800 DSPs. Through this research, we have achieved an FPGA-based accelerator that simultaneously improves the three constraints of inference speed, power consumption, and design cost by using an HLS-based design method and optimization with a slight sacrifice of recognition accuracy.

備考 ： 論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。
Note : Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800 Words (English).

注意：論文要旨は、東工大リサーチリポジトリ(T2R2)にてインターネット公表されますので、公表可能な範囲の内容で作成してください。
Attention: Thesis Summary will be published on Tokyo Tech Research Repository Website (T2R2).