論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Neural Network Hardware Codesign via Nimble FPGA Implementations and Synchronized Spiking in Edge-AI
著者(和文)	Bartels Jim
Author(English)	Jim Bartels
出典(和文)	学位:博士(学術), 学位授与機関:東京工業大学, 報告番号:甲第12922号, 授与年月日:2024年9月20日, 学位の種別:課程博士, 審査員:伊藤 浩之,岡田 健一,德田 崇,原 祐子,白根 篤史,Indiveri Giacomo
Citation(English)	Degree:Doctor (Academic), Conferring organization: Tokyo Institute of Technology, Report number:甲第12922号, Conferred date:2024/9/20, 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	電気電子 電気電子	系 コース	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of	(Philosophy)
学生氏名： Student's Name	Bartels Jim		審査員主査： Chief Examiner	伊藤浩之	

要旨（英文 800 語程度）

Thesis Summary (approx.800 English Words)

This thesis is titled "Neural Network Hardware Codesign via Nimble FPGA Implementations and Synchronized Spiking in Edge-AI" and consists of 9 Chapters.

Chapter 1, "Introduction," provides an overview of the Internet of Things (IoT) and Artificial Intelligence (AI), emphasizing the role and significance of Edge-AI. It sets forth the objective to enhance Edge-AI in terms of capability, accessibility, and efficiency. To achieve this, two distinct hardware codesign strategies are introduced. The first, an application-oriented approach, involves the design of Recurrent Neural Networks (RNNs) and Decision Trees (DTs) on FPGAs for estimating cow behavior and other applications using low-frequency sensors in Edge-AI, detailed in Part I, Chapters 2-5. The second, a dynamics-oriented approach, focuses on the use of synchronization in designing Spiking Neural Networks (SNNs) for FPGAs and a mixed-signal neuromorphic processor, explored in Part II, Chapters 6-8.

Chapter 2, "Edge-AI for Precision Livestock Farming: A Study on Cow Behavior Estimation", outlines the potential of Precision Livestock Farming (PLF) and constraints using AI, followed by two approaches, i.e., the low-power design of a DT and RNN on FPGA for cow behavior monitoring that composes the remainder of Part I.

Chapter 3, "A 216 µW, 87% Accurate Cow Behavior Classifying Decision Tree on FPGA With Interpolated Arctan2", describes the implementation of a Decision Tree (DT) on FPGA, using the cow's neck orientation as a feature to classify cow behavior leading to an accuracy of 86.8%, a power consumption of 216 µW, and an energy consumption of 557 µJ.

Chapter 4, "TinyCowNet: Memory- and Power-Minimized RNNs Implementable on Tiny Edge Devices for Lifelong Cow Behavior Distribution Estimation", proposes a post-training quantization scheme for Recurrent Neural Networks (RNN) and a random search to optimize RNN architectures for cow behavior estimation, achieving high accuracy (<95.5%) with minimal memory (2.043 kB) and operational requirements (42,822 MACs).

Chapter 5, "An Integer-Only Resource-Minimized RNN on FPGA for Low-Frequency Sensors in Edge-AI", enhances the integer-only quantization scheme for RNNs, presenting a resource-minimized hardware architecture with time-and-layer-multiplexing and a custom processing engine for integer-based Multiply-Accumulate (MAC) operations, achieving power consumption between 340 µW and 3.81 mW and energy consumption between 542 and 44.3 µJ, demonstrating the effectiveness of codesigning RNNs in terms of hardware and software for advancing Edge-AI in terms of accessibility and efficiency.

Chapter 6, "An Experimental Study of Synchronization Phenomena in Chaotic Spiking Oscillators Towards Physical Reservoirs", describes an experimental study of synchronization phenomena in a ring network of dual-transistor spiking oscillators, examining the effects of varying supply voltage, coupling strength, and noise, and proposes a physical reservoir utilizing synchronization, setting the stage for the remainder of Part II.

Chapter 7, "A Seizure-Encoding Spiking Neural Network Using Partial Synchronization on a Mixed-Signal Neuromorphic Processor", presents a proof-of-concept of using synchronization to encode seizures from intracranial EEG data with a neural network of Exponential Leaky-Integrate-and-Fire neurons on DYNAP-SE2, highlighting the importance of codesigning SNNs by considering the inherent dynamics on mixed-signal hardware for computation, contributing to the capabilities of Edge-AI.

Chapter 8, "NimbleSNN: A 1.34 µJ/Image 4639 Logic Element Single-Spike Accelerator

on FPGA Using Synchronization", employs Time-To-First Spike (TTFS) coding in SNNs and presents the "NimbleSNN" hardware architecture that employs synchronization in neurons to reduce the number of weights by 49.4% and on-hardware spike buffer size by 78.6%. As a result, NimbleSNN achieves a 20.4 mW power consumption, 65.8 µs inference latency, and 1.34 µJ energy per MNIST image on the ICE40UP5K FPGA while consuming only 4639 logic elements, showcasing a blueprint for SNN accelerators using dynamics-based codesign.

Chapter 9, "Conclusion, Discussion and Future Work", discusses the advancements proposed in the above Chapters and introduces four claims that enhance Edge-AI in terms of capability, accessibility and efficiency by employing the codesign approaches of Part I and II.

In conclusion, this thesis contributes significantly to academics and industry due to the proposed advancements in Edge-AI realized by neural network hardware codesign by means of nimble FPGA implementations and synchronized spiking for computation and optimization showcased on cow behavior estimation and epileptic seizure encoding.