論文 / 著書情報
Article / Book Information

題目(和文)	環境適応型エッジAIの実現に向けた低電力深層学習アクセラレータ
Title(English)	Low-Power Deep Learning Accelerator for Environment-Adaptive Edge AI
著者(和文)	鈴木淳之介
Author(English)	Junnosuke Suzuki
出典(和文)	学位:博士(工学), 学位授与機関:東京科学大学, 報告番号:甲第286号, 授与年月日:2025年3月26日, 学位の種別:課程博士, 審査員:本村 真人,髙橋 篤司,CHU VAN THIEM,ISLAM A K M MAHFUZUL,佐々木 広
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Institute of Science Tokyo, Report number:甲第286号, Conferred date:2025/3/26, 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 情報通信	系 コース	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of	（工学）
学生氏名： Student's Name	鈴木 淳之介		審査員主査： Chief Examiner	本村 真人	

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

Edge AI has emerged as a promising solution to the computational burden challenges of centralized data centers and resource-constrained endpoint devices by enabling data processing directly within endpoint devices. By leveraging improved network traffic, model compression techniques, and the design of domain-specific accelerators, edge AI provides various advantages while operating under strict power and computation constraints, such as low latency, reduced power consumption, and enhanced security.

However, despite these benefits, edge AI systems face the following three fundamental challenges: 1) Power consumption: Operating under strict power and resource constraints in edge devices. 2) Adaptability: Adjusting to temporally and spatially varying computational resource demands and flexibly managing computational cost. 3) Efficiency: Optimizing the tradeoff between computational complexity and inference accuracy within limited computational resources. Limited on-device resources and power constraints often lead to high inference latency and excessive energy usage, resulting in inefficient utilization of computational resources and reduced energy efficiency. Furthermore, since edge AI systems are often optimized for specific tasks and models, they struggle to adapt to dynamic computational demands in diverse environments.

Adaptive inference offers a potential solution to improve the efficiency of edge AI by dynamically adjusting computational resources according to real-time conditions, optimizing the tradeoff between performance and energy consumption. Despite its promise, adaptive inference faces significant challenges in edge AI scenarios. Existing methods often rely on modifying the network architecture, holding multiple model versions, or adopting module-switching mechanisms. These approaches lead to increased memory usage, system complexity, and training costs. To overcome these limitations, this dissertation focuses on a quantization-based method that enables dynamic adaptation using a single model. This approach eliminates the need for architectural modifications and reduces memory overhead, making it suitable for resource-constrained environments.

At the core of this framework is ProgressiveNN, a novel adaptive inference method using bitwise binary (BWB) quantization. BWB quantization enables variable precision by accumulating from the most significant bit (MSB) to the least significant bit (LSB). Desired bitwidth versions are achieved using a single weight set by terminating the accumulation process midway since each bit is independent and can be processed sequentially. This mechanism allows ProgressiveNN to dynamically switch bit-precision, providing flexibility without requiring architectural changes. This consistent and adaptive quantization model is obtained with a single full-parameter training process. The accuracy degradation typically associated with reduced bitwidth is mitigated by retraining unique batch normalization (BN) layers for each bitwidth variation. To maximize the performance of ProgressiveNN, defining an effective method for switching bit precision is crucial. A confidence-based dynamic bit-precision adjustment mechanism is proposed, and it is shown that a threshold defined using entropy effectively balances the accuracy-computation tradeoff. Evaluations on the CIFAR-100 dataset demonstrated a 1.3% improvement in accuracy with an average bitwidth of 2-bit.

Based on ProgressiveNN, this dissertation introduces Pianissimo, an ultra-low-power deep neural network accelerator designed for adaptive inference in extreme edge environments. Pianissimo addresses the severe challenges specific to extreme edge devices, such as limited memory budgets and strict power constraints, while also meeting the need for operational flexibility for varying computational demands. It achieves ultra-low power consumption by simplifying arithmetic units and minimizing data movement through a three-layer memory hierarchy. Additionally, a Block Skip (BS) mechanism designed for integration with event-driven image sensors is proposed. BS significantly reduces computational complexity by dynamically processing only the regions of interest identified by the sensors. Key components enabling these adaptive computations include 1) a progressive MSB

to LSB bit-serial datapath that maximizes arithmetic unit utilization while providing the required bit-precision and 2) a cooperative control mechanism that combines a reduced instruction set computer (RISC) processor with hardware counters. The RISC processor minimizes power consumption by remaining clock-inactive for most of its operating time; thus, the entire control unit consumes only 3.4% of total power. A fabricated 40 nm chip achieves 0.49-1.25 TOPS/W at 0.7 V with MobileNetV1 while supporting a wide range of the latest TinyML models. Furthermore, BS improved the efficiency of convolutional layers from 3.0 TOPS/W to 27.7 TOPS/W, demonstrating Pianissimo's potential for seamless integration with edge sensors.

Furthermore, this dissertation addresses a crucial limitation of the ProgressiveNN model: the lack of zero representation. The ProgressiveNN framework is extended by incorporating Booth encoding to support sparse domains. Additionally, ProgressiveNN initially faced significant accuracy degradation at low bitwidths. To overcome this, a ternary distribution pre-fixing training strategy is introduced, enabling the creation of a consistent single model while minimizing accuracy loss. Evaluations using ResNet18 on the ImageNet dataset demonstrated negligible accuracy degradation when transitioning from a 61.4% ternary representation to an 8-bit configuration. These results suggest the potential for further efficiency improvements by utilizing sparse networks and bit-level sparse datapaths.

This dissertation overcomes the limitations of conventional fixed-edge AI through a comprehensive proposal spanning from adaptive algorithms to chip-level implementation. The main contributions of this dissertation are as follows: 1) Progressive bit-scalable networks that enable flexible tradeoffs between computational complexity and inference accuracy using a single model with a bit of additional training, 2) a confidence-based dynamic bit-precision adjustment strategy supported by a specialized dataflow design that achieves computational scalability, energy efficiency, and fine-grained precision control in resource-constrained environments, and 3) an adaptive accelerator design featuring a progressive bit-serial datapath, efficient memory hierarchy, and SW-HW cooperative control, enabling sub-mW power consumption and scalable ultra-low-power edge AI systems. These contributions establish a pathway toward realizing dynamic, efficient, and scalable AI systems capable of adapting to diverse and resource-constrained environments.