論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Algorithm-architecture Collaborative Research toward Beyond-CNNs
著者(和文)	YANJIALE
Author(English)	Jiale Yan
出典(和文)	学位:博士(学術), 学位授与機関:東京科学大学, 報告番号:甲第24号, 授与年月日:2024年12月31日, 学位の種別:課程博士, 審査員:本村 真人,一色 剛,髙橋 篤司,原 祐子,藤木 大地
Citation(English)	Degree:Doctor (Academic), Conferring organization: Institute of Science Tokyo, Report number:甲第24号, Conferred date:2024/12/31, 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 (学術) Doctor of
学生氏名： Student's Name	JIALE YAN	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 ）

This dissertation introduces algorithm-architecture collaborative designs to reduce memory demands and improve the edge device's performance for Beyond-CNNs. The contributions of this thesis include the following:

1. An efficient parameter reduction method, tensor-train decomposition (TTD), is explored and implemented within Vision-Multilayer perceptron (MLP). The thesis provides an in-depth discussion of the implementation details and algorithmic optimizations of TTD in fully connected (FC) layers, presenting an efficient approach to reducing model size.
2. A hardware-aware lottery ticket hypothesis (LTH) framework is explored, integrating both pruning and quantization without compromising accuracy when applied to graph neural networks (GNNs). This unified framework extends the conventional LTH from CNNs to GNNs, enabling significant memory reduction while maintaining model performance.
3. Systemic and practical examples for beyond-CNN architectures are explored, focusing on energy efficiency, performance, and scalable extensions. The dissertation explores the efficient dataflow, mapping, and tiling methods for implementing different computation kernels, such as convolution (CONV) and deconvolution (DeCONV) layers, including SpMM and GeMM. Finally, it proposes efficient architecture designs for acceleration, especially for generative networks (GANs) and graph neural networks (GNNs). It provides insights into the hardware-software co-design necessary for efficient AI model deployment.

From the algorithm side, we explore TTD as a solution to reduce the parameter size of Vision MLP. This method demonstrates how FC layers can be compressed while preserving their learning capabilities. By applying TTD methods, we revealed that different components of Vision MLPs possess different learning capacities and robustness when subjected to decomposition. We also put forward the Train-TTD-Train method, which enhancs the trade-off between model accuracy and size. In the evaluation, the proposed method showed a better trade-off than conventional TTD methods on ImageNet-1K and also achieved a comparable accuracy with a memory reduction on Cifar-10. It confirms the viability of TTD as an essential tool in reducing the footprint of modern AI models.

Additionally, we extended the LTH from CNNs to GNNs. It demonstrated the existence of high-performing subnetworks within randomly initialized models, discoverable through pruning a GNN without any weight training. Based on SLT, we introduced Multicoated Super-masks and implemented them in GNNs by proposing a strategy for setting its pruning threshold adaptively. For deep GNNs, we uncover untrained recurrent networks, which exhibit performance on par with their trained feed-forward counterparts. Through the evaluation of various datasets, including the Open Graph Benchmark, we have established a triple-win scenario for SLTH-based GNNs: by achieving high sparsity, competitive performance, and high memory efficiency, it demonstrates suitability for energy-efficient graph processing.

The dissertation further addresses the architectural challenges of system design. We explored an energy-efficient architecture to support both CONV and DeCONV layers. Since GANs have become ubiquitous in image generation applications, which include CONV layers, CONV-based residual blocks, and DeCONV layers, previous accelerator studies focus too much on optimizing CONV layer computation but have low PE utilization in computing residual blocks and DeCONV layers. To solve these challenges, we propose a dual convolution mapping method for CONV and DeCONV layers to fully utilize the available PE resources. A cross-layer scheduling method is also proposed to avoid extra off-chip memory access in residual block processing. Precision-adaptive PEs are used to support flexible bit-widths for inputs and weights. We implement a generative network accelerator (GNA). The GNA achieves an energy efficiency of 2.05 TOPS/W. These theoretical insights and practical experience also provide foundational support for designing other accelerators.

Compared with CONV and DeCONV, accelerating unregular computation patterns such as graph processing is still challengble. We proposed a co-design framework to tackle the specific challenges of scaling GNNs, focusing on reducing memory communication overhead and addressing irregular memory access patterns. The real-time, large-scale GNN inference faces challenges due to the size of node features and sparse adjacency matrices, leading to memory communication and buffer size overheads. While graph partitioning can help with localization and reduce on-chip buffer size, fine-grained partitioning results in increased inter-partition edges and off-chip memory accesses, negatively impacting overall performance. To overcome these limitations, we proposed GCNBingo, a scalable GNN acceleration framework that

introduces multidimensional dynamic feature summarization called Cross-Partition Message Quantization for inter-partition message passing. This eliminates irregular off-chip memory access without additional training and accuracy loss, even with fine-grained partitioning. We also extended the SLT to reduce the model size and a hardware architecture based on GNA designs to exploit fine-grained sparsity for improved load balancing. Our FPGA-based implementation significantly reduces memory access while maintaining accuracy on par with the original model.

As AI models evolve, large language models (LLMs) are increasingly prevalent, posing new challenges to hardware memory and computation. To address these, researchers may adopt more aggressive compression methods, such as lower bit precision and higher sparsity, while maintaining accuracy. The algorithmic strategies in this dissertation offer potential for LLM optimization, particularly in compressing attention mechanisms and linear layers. Moreover, the hardware architecture solutions provide efficient resource utilization and performance improvements, impacting scalable neural network acceleration techniques.