

OPTIMIZING NEURAL NETWORKS PERFORMANCE ON PARALLEL ARCHITECTURES

SAURABH TEWARI



DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY DELHI

JANUARY 2024

© Saurabh Tewari, Indian Institute of Technology Delhi (IITD), 2024.

All rights reserved.

OPTIMIZING NEURAL NETWORKS PERFORMANCE ON PARALLEL ARCHITECTURES

by

SAURABH TEWARI

Department of Computer Science and Engineering

Submitted

in fulfillment of the requirements of the degree of Doctor of Philosophy

to the



INDIAN INSTITUTE OF TECHNOLOGY DELHI

JANUARY 2024

Certificate

This is to certify that the thesis titled “**OPTIMIZING NEURAL NETWORKS PERFORMANCE ON PARALLEL ARCHITECTURES**” being submitted by **Saurabh Tewari** for the award of **Doctor of Philosophy** in Computer Science and Engineering is a record of bonafide work carried out by him under my guidance and supervision in the Department of Computer Science and Engineering, Indian Institute of Technology Delhi. The work presented in this thesis has not been submitted elsewhere, either in part or full, for the award of any other degree or diploma unless otherwise stated explicitly.

The works presented in Chapter 5 on Self Organizing Maps (SOMs) were conducted in collaboration with students at KTH Royal Institute of Technology, Sweden. We acknowledge their valuable contributions explicitly in the corresponding chapter.

Anshul Kumar

Professor

Department of Computer Science and Engg.

Indian Institute of Technology Delhi

New Delhi- 110016

Kolin Paul

Professor

Department of Computer Science and Engg.

Indian Institute of Technology Delhi

New Delhi- 110016

Acknowledgments

I am deeply grateful to my esteemed Ph.D. supervisors, Prof. Anshul Kumar and Prof. Kolin Paul, for their exceptional guidance and unwavering support throughout my research journey. Their technical expertise and invaluable insights have been instrumental in shaping my academic growth and thesis work. They have provided mentorship and valuable advice on various aspects of academic life, proving to be excellent mentors in every sense. Their constant availability for discussions, constructive suggestions, and motivation to complete my thesis on time have been indispensable. During the challenging times of the COVID-19 lockdown, their unwavering support kept me focused and encouraged.

Heartfelt appreciation goes to my research committee members, Prof. M. Balakrishnan and Prof. Preeti Ranjan Panda, for their valuable suggestions and feedback on my thesis work, enriching its quality and depth. I am also sincerely grateful to Prof. Ahmed Hemani, a professor at KTH Sweden, for allowing me to work in their state-of-the-art CGRA accelerator labs, significantly enhancing my research exposure and knowledge.

I extend my gratitude to all the instructors at IIT Delhi who imparted valuable skills during my coursework, contributing significantly to my overall academic growth.

Special thanks go to my friends at IIT Delhi, Rajesh Kedia, Sandeep Kumar, Dishant Goyal, and Omais Shafi, for their unwavering support and valuable suggestions on useful writing tools, aiding me in crafting research papers and creating graphical representations. In particular, I want to acknowledge Rajesh Kedia for his meticulous review and invaluable feedback. I am grateful to Vandana Ahluwalia, the lab in-charge, for providing access to the labs and resources and to Som Dutt Sharma from the DHD-LAB for granting access to the FPGA boards and essential software, which played a critical role in my research.

I am deeply thankful to my family members, including my wife, parents, and daughters, whose unwavering support and understanding made this achievement possible. Their love, encouragement, and sacrifices have been a constant source of motivation and strength. Lastly,

I express gratitude to the Almighty for giving me the strength and opportunity to work with wonderful and helpful people.

I also want to acknowledge all those whose names may not be mentioned here but have contributed in various ways, big or small, to my personal and academic growth. Their collective support and encouragement have been the driving force behind the successful completion of my Ph.D. thesis, and I humbly acknowledge their invaluable contributions.

Saurabh Tewari

Abstract

In recent years, Neural Networks (NNs) have experienced tremendous growth in applications across diverse fields, revolutionizing industries such as healthcare, agriculture, surveillance, and autonomous systems. The availability of large datasets and powerful computing systems, coupled with user-friendly libraries like Tensorflow and PyTorch, has enabled NNs to achieve human-like performance. As a result, NN-based applications have become an integral part of our daily lives, from face recognition and chatbots to shopping recommendations and medical diagnoses.

This Ph.D. thesis focuses on optimizing the inferencing phase of NNs for edge devices with limited computing resources and tight energy budgets. Edge devices, such as smartphones and wearable devices, are now widely equipped with NN applications, necessitating energy-efficient solutions for improved user experiences and privacy protection. The study encompasses three classes of NNs: Self Organizing Maps (SOMs), Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs), each posing unique challenges in terms of data reuse and energy efficiency.

The thesis introduces the neural network model, illustrating the structure and different types of NNs, including feedforward, recurrent, and deep neural networks. The development and deployment phases of NN applications are also explained, emphasizing the significance of optimizing the inferencing phase for edge devices.

In the context of CNNs, the research focuses on optimizing data reuse through partitioning and scheduling schemes. An analytical framework is developed to estimate off-chip memory

accesses, considering architectural constraints and data shape. The framework enables the comparison of various partitioning and scheduling approaches, facilitating the discovery of optimal solutions to improve energy efficiency and throughput for different CNN layers.

For RNNs, which present additional challenges due to their dependency on previous time-step computations, a novel data reuse approach is proposed. This approach efficiently reuses weights of large matrices for consecutive time steps, irrespective of on-chip memory size. FPGA implementations of Long-Short Term Memory Network (LSTM) accelerators demonstrate the effectiveness of this approach in enhancing energy efficiency and throughput for popular LSTM models.

For SOMs, the study explores the impact of quantization techniques on accuracy and energy efficiency. A custom semi-systolic array design is devised, and a trade-off between NN accuracy and energy consumption is analyzed for different bit-width implementations. The study analyzes the benefits and trade-offs of using different bit resolutions for SOMs, catering to energy-constrained systems where area, power, and performance are critical considerations.

In conclusion, this Ph.D. thesis advances the state-of-the-art energy-efficient acceleration of modern NNs for edge devices. By proposing novel data reuse techniques and analytical frameworks, the research provides valuable insights for improving the inferencing phase, thus contributing to the widespread adoption of NN-based applications in energy-constrained environments. The work also opens new avenues for future research, pushing the boundaries of NN optimization and edge AI applications.

सार

हाल के वर्षों में, न्यूरोल नेटवर्क्स (एनएन) ने विभिन्न क्षेत्रों में अनुप्रयोगों में जबरदस्त वृद्धि का अनुभव किया है, जिससे स्वास्थ्य देखभाल, कृषि, निगरानी और स्वायत्त प्रणालियों जैसे उद्योगों में क्रांति आ गई है। बड़े डेटासेट और शक्तिशाली कंप्यूटिंग सिस्टम की उपलब्धता, साथ में टेन्सरफ्लो और पायटोरच जैसी उपयोगकर्ता-अनुकूल लाइब्रेरीज़ ने एनएन को मानव-जैसा प्रदर्शन प्राप्त करने में सक्षम बनाया है। परिणामस्वरूप, एनएन-आधारित एप्लिकेशन चेहरे की पहचान और चैटबॉट से लेकर शॉपिंग अनुशंसाओं और चिकित्सा निदान तक हमारे दैनिक जीवन का एक अभिन्न अंग बन गए हैं।

यह पीएच.डी. थीसिस सीमित कंप्यूटिंग संसाधनों और तंग ऊर्जा बजट वाले एज उपकरणों के लिए न्यूरोल नेटवर्क (एनएन) के अनुमान चरण को अनुकूलित करने पर केंद्रित है। स्मार्टफोन और पहनने योग्य डिवाइस जैसे एज डिवाइस अब व्यापक रूप से एनएन अनुप्रयोगों से सुसज्जित हैं, जिससे बेहतर उपयोगकर्ता अनुभव और गोपनीयता सुरक्षा के लिए ऊर्जा-कुशल समाधान की आवश्यकता होती है। अध्ययन में एनएन के तीन वर्ग शामिल हैं: सेल्फ ऑर्गनाइजिंग मैप्स (एसओएम), कन्वेंशनल न्यूरोल नेटवर्क्स (सीएनएन), और रिकरंट न्यूरोल नेटवर्क्स (आरएनएन), प्रत्येक वर्ग डेटा पुनः उपयोग और ऊर्जा दक्षता के मामले में अद्वितीय चुनौतियां पेश करता है।

थीसिस न्यूरोल नेटवर्क मॉडल को पेश करने से शुरू होती है, जिसमें फीडफॉरवर्ड, रिकरंट और डीप न्यूरोल नेटवर्क सहित संरचना और विभिन्न प्रकार के एनएन को दर्शाया गया है। एनएन अनुप्रयोगों के विकास और तैनाती चरणों को भी समझाया गया है, जिसमें एज उपकरणों के लिए अनुमान चरण को अनुकूलित करने के महत्व पर जोर दिया गया है।

सीएनएन के संदर्भ में, अनुसंधान विभाजन और शेड्यूलिंग योजनाओं के माध्यम से डेटा के पुनः उपयोग को अनुकूलित करने पर केंद्रित है। सिस्टम-आर्किटेक्चरल बाधाओं और डेटा आकार पर विचार करते हुए, ऑफ-चिप मेमोरी एक्सेस का अनुमान लगाने के लिए एक विश्लेषणात्मक ढांचा विकसित किया गया है। ढांचा विभिन्न विभाजन और शेड्यूलिंग दृष्टिकोणों की तुलना करने में सक्षम बनाता है, जिससे विभिन्न सीएनएन परतों के लिए ऊर्जा दक्षता और थ्रूपुट में सुधार के लिए इष्टतम समाधान की खोज की सुविधा मिलती है।

आरएनएन के लिए एक नया डेटा पुनः उपयोग दृष्टिकोण प्रस्तावित है, जो पिछले समय-चरण गणनाओं पर निर्भरता के कारण अतिरिक्त चुनौतियां पेश करता है। यह कार्य ऑन-चिप मेमोरी क्षमता के बावजूद, किसी भी दो सन्निहित समय चरणों के लिए बड़े मैट्रिक्स के वजन का कुशलतापूर्वक पुनः उपयोग करता है। लॉन्ग-शॉर्ट टर्म मेमोरी नेटवर्क (एलएसटीएम) एक्सेलेरेटर का एफपीजीए कार्यान्वयन लोकप्रिय एलएसटीएम मॉडल के लिए ऊर्जा दक्षता और थ्रूपुट बढ़ाने में इस दृष्टिकोण की प्रभावशीलता को प्रदर्शित करता है।

एसओएम के लिए, अध्ययन सटीकता और ऊर्जा दक्षता पर परिमाणीकरण तकनीकों के प्रभाव का पता लगाता है। एक कस्टम सेमी-सिस्टोलिक सरणी डिज़ाइन तैयार किया गया है, और विभिन्न बिट रिज़ॉल्यूशन कार्यान्वयन के लिए एनएन सटीकता और ऊर्जा खपत के बीच एक व्यापार-बंद का विश्लेषण किया गया है। अध्ययन एसओएम के लिए अलग-अलग बिट रिज़ॉल्यूशन का उपयोग करने के लाभों और व्यापार-बंदों का विश्लेषण करता है, ऊर्जा-सीमित प्रणालियों को पूरा करता है जहां क्षेत्र, शक्ति और प्रदर्शन महत्वपूर्ण विचार हैं।

अंत में, यह पीएच.डी. थीसिस एज उपकरणों के लिए आधुनिक एनएन के ऊर्जा-कुशल त्वरण में अत्याधुनिक को आगे बढ़ाती है। उपन्यास डेटा पुनः उपयोग तकनीकों और विश्लेषणात्मक ढांचे का प्रस्ताव करके, यह शोध अनुमान चरण में सुधार के लिए मूल्यवान अंतर्दृष्टि प्रदान करता है, इस प्रकार ऊर्जा-सीमित वातावरण में एनएन-आधारित अनुप्रयोगों को व्यापक रूप से अपनाने में योगदान देता है। यह कार्य एनएन अनुकूलन और एज एआई अनुप्रयोगों की सीमाओं को आगे बढ़ाते हुए भविष्य के अनुसंधान के लिए नए रास्ते भी खोलता है।

Contents

Certificate	i
Acknowledgments	iii
Abstract	v
List of Figures	xiii
List of Tables	xv
1 Introduction	1
1.1 Neural Network Model	2
1.2 Phases of NN Applications: Development to Deployment	3
1.3 Edge AI: Inferencing on Edge Devices	4
1.4 Energy Efficient NN Inferencing	6
1.4.1 Previous Work	6
1.4.2 Thesis objectives and contributions	9
1.5 Summary and Outline of the Thesis	11
2 Analyzing Off-chip Memory Accesses	13
2.1 Analytical study of architectural parameters	18
2.1.1 Aligned Access	20

2.1.2	General Case	20
2.2	Off-Chip Memory Access of 3D Data	21
2.3	Implementation and Results	25
2.3.1	Results	26
2.4	Summary	27
3	Optimizing the Performance of CNN Accelerators	29
3.1	Related Work	34
3.2	Off-Chip Memory Accesses of CNN Layers	35
3.2.1	Trip Counts of Tiles	35
3.2.2	Off-chip memory access of CL	39
3.2.3	Optimization problem	40
3.2.4	Off-chip memory access of FCL	40
3.3	Implementation and Results	41
3.3.1	Implementation	41
3.3.2	Validation	41
3.3.3	Benchmarks	42
3.3.4	Baselines	42
3.3.5	Results	43
3.4	Summary	47
4	Optimizing the Performance of RNN/LSTM Accelerators	49
4.1	Introduction	49
4.2	Background	50
4.3	Related Work	52
4.4	Split And Combine Computations Approach	54
4.4.1	Basic Approach	54
4.4.2	Block-wise reuse	57

4.5	Experimental Setup And Results	62
4.5.1	Baseline	64
4.5.2	Benchmarks	64
4.5.3	Results	65
4.6	Summary	69
5	Performance Improvement of SOM by using Low Bit-Width Resolution	71
5.1	Introduction	72
5.2	Background	74
5.3	Low bit-width FPGA Design of SOM	76
5.4	Experimental Results And Analysis	80
5.4.1	Accuracy experimental setup	80
5.4.2	FPGA experimental setup	82
5.5	Summary	84
6	Conclusion And Future Directions	87
6.1	Introduction	87
6.2	Recapitulation of Research Objectives	88
6.3	Summary of Key Findings	88
6.4	Implications and Future Research Directions	89
6.5	Closing Remarks	90
6.6	Conclusion	90
	Bibliography	93
	List of Publications	101
	Biography	103

List of Figures

1.1	Example of simple Neural Network.	2
1.2	Work Flow for Neural Networks.	3
1.3	Key challenges for Edge-AI Accelerator	5
1.4	Typical DNN accelerator architecture.	6
1.5	Broad Classification of previous works for improving the performance of DNN accelerators.	7
2.1	Memory accesses and energy estimates for accessing data (a) always from the off-chip memory. (b) from two-level of memory hierarchy while reusing the data from on-chip memory.	14
2.2	Read/Write of inputs, weights, and partial/final outputs from different levels of memories.	15
2.3	Impact of tile dimensions on off-chip memory accesses (a) off-chip memory access variations of a single convolution layer when accessing with tiles of different dimensions, x-axis represents a subset of valid tile space of the layer (b) Impact of good and bad tile dimensions on different convolution layers of VGG16.	16
2.4	Analytical framework to estimate the performance, off-chip memory accesses and energy of DNNs.	17
2.5	A 2D data and memory accesses on 64-bit data bus	18

2.6	Off-chip memory accesses on 64-bit wide data bus	19
2.7	Off-Chip Accesses and a 3D partitioned data and tiles	21
2.8	Off-chip memory accesss of VGG16 layers (a) Comparison between the off-chip memory accesses estimated by the analytical framework and measured on FPGA. (b) Breakdown of off-chip memory access by data type estimated by the analytical framework.	26
3.1	Convolution and fully connected layers	30
3.2	Input and Output dimensions of CL layer for filter stride=1	31
3.3	CNN layer tiles in off-chip and on-chip memory	32
3.4	Parameters and activations proportions in CL and FC layers of CNNs.	34
3.5	A Convolution layer data partitioned into tiles	36
3.6	Different Scheduling of tiles.	36
3.7	Off-chip memory access of convolution layers for 8 and 16 bits data width. BWA: Bus Width Aware, SS: SmartShuttle	43
3.8	Layer wise off-chip memory access for IRO, ORO and WRO schemes.	44
3.9	Off-chip memory access for varying on-chip buffer sizes. BWA: Bus Width Aware, SS:SmartShuttle	45
3.10	Off-chip memory access of Fully connected layers. BWA: Bus Width Aware, SS:SmartShuttle	45
3.11	Energy and latency efficiency. BWA: Bus Width Aware, SS:SmartShuttle . . .	46
4.1	Data dependency in computations of LSTMs between consecutive time steps .	51
4.2	Proposed approach: Splitting computations into partial sums and reusing the weights of R	52
4.3	Fraction of partial sums computed using (a) lower diagonal element of R (b) upper diagonal elements of R	55
4.4	Computation stages at time step t	56

4.5	Computation stages at time step $t+1$	57
4.6	(a) Partitions of a matrix $R_{N \times N}$ into $B \times B$ square matrices. (b) $R^L = \{P_{(r,m)} : r \geq m\}$ (b) $R^U = \{P_{(r,m)} : r < m\}$	58
4.7	Partitions of R^L and R^U required for computations of partial sum S^L and S^U	59
4.8	Block level FPGA Design for LSTM Implementation.	63
4.9	Off-chip memory access for matrix-vector multiplication of two consecutive time steps with different on-chip buffer to R matrix size ratio. (a) using analytical framework (b) measured on hardware	65
4.10	Throughput variation with different on-chip buffer/R ratio	66
4.11	Throughput variation with compute resources for 50% on-chip buffer/R ratio	67
4.12	Energy improvement with on-chip buffer size/R	68
5.1	SOM based Genomic Identification	74
5.2	High level block diagram of the FPGA Implementation of BioSOM.	77
5.3	Pipelined implementation of Compute_Distance with M stages and Initiation Interval (II)=1	79
5.4	Pipelined Implementation of Neuron_Update with 7 stages and Initiation Interval (II)=1.	80
5.5	Quantization error compared to the identification error.	82
5.6	Area and energy comparison for different fixed-point format.	83

List of Tables

2.1	Off-Chip memory accesses of the tiles of size 5×5	19
4.1	FPGA resource utilization for $B=64$, $PARL_FACTOR=8$, and $N=128$	64
4.2	LSTM Models used for experiments	64
5.1	Resource Comparison of different fixed point formats	83

