OPTIMIZING NEURAL NETWORKS PERFORMANCE ON PARALLEL ARCHITECTURES

SAURABH TEWARI

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY DELHI

JANUARY 2024

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, न्यूरल 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 टेन्सरफ़्लो 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 न्यूरल नेटवर्क (एनएन) के अनुमान चरण को अनुकूलित करने पर केंद्रित है। स्मार्टफ़ोन 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), कन्वेन्शनल न्यूरल नेटवर्क्स (सीएनएन), और रिकरंट न्यूरल नेटवर्क्स (आरएनएन), प्रत्येक वर्ग डेटा पुन: उपयोग और ऊर्जा दक्षता के मामले में अद्वितीय चुनौतियां पेश करता है।

थीसिस न्यूरल नेटवर्क मॉडल को पेश करने से शुरू होती है, जिसमें फीडफॉरवर्ड, रिकरंट और डीप न्यूरल 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 पुन: 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 पुन: 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 पुन: 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 पुन: 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

List of Figures

List of Tables