EXECUTION TIME PREDICTION FOR CONCURRENT CNNS ON FPGA ACCELERATORS

SHIKHA GOEL

THE AMAR NATH AND SHASHI KHOSLA SCHOOL OF INFORMATION TECHNOLOGY
INDIAN INSTITUTE OF TECHNOLOGY DELHI
MAY 2024

EXECUTION TIME PREDICTION FOR CONCURRENT CNNS ON FPGA ACCELERATORS

by

SHIKHA GOEL

The Amar Nath and Shashi Khosla School of Information Technology

Submitted

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

to the

INDIAN INSTITUTE OF TECHNOLOGY DELHI

MAY 2024

DEDICATED TO

My mother Meena Goel.

Certificate

This is to certify that the thesis titled "**Execution time prediction for concurrent CNNs on FPGA accelerators**" being submitted by **Shikha Goel** for the award of **Doctor of Philosophy** in Information Technology is a record of bonafide work carried out by her under my guidance and supervision in the Amar Nath and Shashi Khosla School of Information Technology, 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. Certain work included in this thesis involved collaboration which has been explicitly specified/acknowledged in the corresponding chapter and the part done by the collaborator appeared in his respective reports/thesis.

M. Balakrishnan
Professor
Department of Computer Science and Engg.
Indian Institute of Technology Delhi
New Delhi- 110016

Rijurekha Sen
Professor
Department of Computer Science and Engg.
Indian Institute of Technology Delhi
New Delhi- 110016

Acknowledgments

"Kind words can be short and easy to speak, but their echoes are truly endless."

—Mother Teresa

My Ph.D. journey was a complete roller coaster ride. As I look back, I can say with confidence that it was an incredibly valuable experience. It brought about a profound transformation in my personality, my thought processes, and my overall character. It has taught me a lot of patience. I am deeply grateful to all the individuals who have played a part in my journey, and I will make every effort to acknowledge their contributions.

First and foremost, I want to extend my sincere appreciation to **Prof. M. Balakrishnan**, my supervisor. He is not just a mentor but a true believer in my potential. It was his faith in my abilities that led me from the M.Tech program to a Ph.D. His work in the field of aiding visually impaired individuals has always been a great inspiration. Despite his busy schedule, he has been incredibly supportive, providing guidance and being available for discussions. His dedication and unwavering support has truly motivated me. I also want to acknowledge my co-supervisor, **Prof. Rijurekha Sen**. Our journey began in her first year at IIT Delhi, which coincided with my first year of Ph.D. I enrolled in her machine learning course, the very first one she taught at IIT Delhi. Throughout my academic path, she has provided practical insights and valuable input into my thesis work. Beyond academics, Riju has always made herself available for discussions on a wide range of topics, from career aspirations to casual conversations. She has been an exceptional source of mental support and encouragement. Both Prof. Balakrishnan and Prof. Sen have been pillars of support during challenging phases of my research, particularly during the COVID-19 pandemic. Their unwavering motivation and encouragement have been pivotal in my academic journey, and for that, I am immensely grateful.

I would like to extend my heartfelt thanks to Rajesh Kedia for his support and mentorship throughout my Ph.D. journey. His guidance and teachings have enriched my understanding in numerous ways. Rajesh was always available for discussions and guidance, regardless of the hour. Whether it was a query at 7 in the morning or 10 at night, he was there to provide assistance. He patiently clarified even the smallest of doubts and played a crucial role in deepening my knowledge of FPGAs. I cannot thank him enough, and I am profoundly grateful for having met him during my Ph.D. I also wish to express my gratitude to Prof. Anshul Kumar, Prof. Kolin Paul, and Prof. Preeti R. Panda for their invaluable feedback on my research. I had the privilege of being a teaching assistant for some of Prof. Anshul Kumar's courses, which proved to be a tremendous learning experience. My heartfelt appreciation also

goes out to my close friends during my Ph.D., namely, Dilpreet Kaur, Ovia Seshadri, Sandeep Kumar, and Arindam Bhattacharya. My Ph.D. journey was made enjoyable thanks to the wonderful moments we shared, from tea breaks to lunches. Our friendship extends beyond our research, and we have stayed in touch, catching up whenever possible. I would like to acknowledge all my lab mates, who made me feel like I was never alone in the lab. Additionally, I extend my gratitude to the lab and office administrators who have always been helpful in meeting various logistical requirements.

My Ph.D. journey has been filled with numerous highs and lows. Throughout this journey, the support of my family has been my strength. I am forever grateful to them for their immense contributions. First and foremost, I want to express my deepest appreciation to my parents, who have played a pivotal role in making me capable of undertaking this journey. I would like to give special thanks to my mother, Meena Goel, whose belief in me has been a constant source of inspiration. I dedicate this Ph.D. to her. Every milestone I have achieved is a testament to her sacrifices and endless support. I would like to express my deep gratitude to my husband, Vishesh Garg, who has been my unwavering pillar of strength throughout this journey. From the moment I decided to pursue my Ph.D. until the very end of it, he has been a constant source of motivation. During times when I completely lost confidence in myself, he always believed in me and instilled a tremendous amount of confidence. I am truly grateful that we made the decision to get married during my Ph.D. With his continuous support, my Ph.D. journey became incredibly smooth, even after our marriage.

I would like to express my gratitude to my brother Sahil Goel, and sister-in-law Megha Goel and my adorable nephew Hitansh Goel, who played a significant role in keeping me grounded during challenging times. My brother's trust and unwavering support throughout my Ph.D. journey were invaluable, and my sister-in-law took excellent care of me during difficult moments. My gratitude extends to my in-laws, Usha Garg and Ram Niwas Garg, brother-in-law Vishal Garg, sister-in-law Anu Garg and nephew Vihaan Garg for their love, understanding. A special thanks to my mother-in-law for taking on household responsibilities, enabling me to concentrate on my studies. They are my pillars of strength, providing the comfort and love I needed. I also want to thank my sister Sonal Aggarwal, and nephew, who played a significant role in the final phase of my Ph.D. journey. When we moved to Bangalore during this crucial period, my sister provided the warmth and comfort of home, especially when I missed my mother. I will always be grateful to her for her motherly love, which was a source of immense comfort when I was away from home.

Shikha Goel

Abstract

In recent years, Convolutional Neural Networks (CNNs) have gained immense popularity for their ability to deliver satisfactory accuracy in various machine learning and computer vision applications. The landscape is rich with a wide array of CNNs, each differing significantly in terms of execution time, energy consumption, and inference accuracy. To accelerate CNN execution, several hardware accelerators have been introduced, with Xilinx's Deep Learning Processor Unit (DPU) being a notable example. The DPU offers a multitude of configuration options and is capable of executing a diverse range of CNNs. However, the choice of CNNs and the configuration options in these accelerators create a vast design space, necessitating detailed trade-off analysis to design efficient systems. To support such analysis, this thesis introduces several variations to a framework for estimating the execution time and energy consumption of CNNs when executed on DPU-based systems. This framework offers invaluable assistance to system designers in selecting accelerator and CNN configuration parameters based on specific application requirements, ultimately leading to optimized performance.

Moreover, the methodologies proposed in this thesis are generic and can be applied to a wide range of CNN accelerators. The thesis begins with the introduction of INFER, a methodology designed to estimate the execution time of any CNN on a given DPU size without the need for actual implementation. INFER operates within restricted DPU configurations, allowing up to only three DPUs active concurrently and employing the default bus interconnections provided by the Xilinx tool to the external memory. Additionally, it can estimate the additional time required due to memory interference resulting from the concurrent use of multiple DPUs. Energy estimation for CNNs running on a DPU is another focus of this thesis. An energy estimation technique named EnergyNN, is developed by leveraging the characteristics of CNNs and DPUs. This technique can be applied to predict the energy consumption of even newer CNNs not used in the model development. Extensive evaluation demonstrates the effectiveness of these approaches, with average prediction errors of 6.6% for execution time and 8.8% for energy estimation. The utility of these predictions is further illustrated through real-world applications in traffic monitoring and drone systems.

The thesis also expands its scope to explore the design space consisting of a larger number of DPUs, thus effectively increasing the concurrency. However, this expansion introduces complexities related to bus connections and CPU core limitations. To address these challenges, the thesis introduces an execution time predictor designed to optimize DPU configurations to meet the performance demands of diverse tasks. Various methods are employed to isolate concurrent CNNs from each other and the operating system. A machine learning-based prediction approach named EXPRESS, is then introduced to predict the execution time of a given CNN on a DPU configuration, considering the characteristics of the CNN, DPU, and bus. EXPRESS provides predictions for both CPU and DPU processing times, resulting in the estimation of end-to-end processing times.

To further enhance this approach, EXPRESS-2.0 is introduced, offering support for heterogeneous CNNs. To avoid having different number of features for different count of DPUs/CNNs, EXPRESS-2.0 consolidates the features of Co-runners (CNNs which run concurrently with the CNN for which we predict the execution time) together. A controller has been developed to isolate CPU cores responsible for executing the operating system and the actual CNNs. This isolation reduces the variation in execution time measurements and enhances prediction accuracy. Across all these methodologies, machine learning based regression techniques are employed for prediction. All experiments are conducted on a real FPGA board, and the evaluation, featuring 16 standard CNNs, reveals very low prediction errors. EXPRESS and EXPRESS-2.0 achieve an average prediction error of 2.2% and 0.7%, respectively. The low prediction error of the framework make it highly effective for design space exploration, offering significant benefits to embedded system application developers.

सार

हाल के वर्षों में, कन्वेन्शनल न्यूरल नेटवर्क्स (सीएनएन) ने विभिन्न मशीन लर्निंग और कंप्यूटर विज़न अनुप्रयोगों में संतोषजनक सटीकता प्रदान करने की अपनी क्षमता के लिए काफी लोकप्रियता हासिल की है। परिदृश्य सीएनएन की एक विस्तृत श्रृंखला से समृद्ध है, प्रत्येक निष्पादन समय, ऊर्जा खपत और अनुमान सटीकता के मामले में काफी भिन्न है। सीएनएन निष्पादन में तेजी लाने के लिए, कई हार्डवेयर त्वरक पेश किए गए हैं, जिसमें ज़ाइलिंक्स की डीप लर्निंग प्रोसेसर यूनिट एक उल्लेखनीय उदाहरण है। डीपीयू कई कॉन्फ़िगरेशन विकल्प प्रदान करता है और सीएनएन की एक विविध श्रृंखला को निष्पादित करने में सक्षम है। हालाँकि, सीएनएन की पसंद और इन त्वरक में कॉन्फ़िगरेशन विकल्प एक विशाल डिज़ाइन स्थान बनाते हैं, जिससे कुशल सिस्टम को डिज़ाइन करने के लिए विस्तृत ट्रेड-ऑफ विश्लेषण की आवश्यकता होती है। इस तरह के विश्लेषण का समर्थन करने के लिए, यह थीसिस डीपीयू-आधारित सिस्टम पर निष्पादित होने पर सीएनएन के निष्पादन समय और ऊर्जा खपत का अनुमान लगाने के लिए एक रूपरेखा में कई बदलाव पेश करती है। यह ढांचा विशिष्ट एप्लिकेशन आवश्यकताओं के आधार पर एक्सेलेरेटर और सीएनएन कॉन्फ़िगरेशन पैरामीटर का चयन करने में सिस्टम डिजाइनरों को अमूल्य सहायता प्रदान करता है, जिससे अंततः अनुकूलित प्रदर्शन होता है।

इसके अलावा, इस थीसिस में प्रस्तावित पद्धतियाँ सामान्य हैं और इन्हें सीएनएन त्वरक की एक विस्तृत श्रृंखला पर लागू किया जा सकता है। थीसिस इन्फर की शुरूआत के साथ शुरू होती है, जो वास्तविक कार्यान्वयन की आवश्यकता के बिना किसी दिए गए डीपीयू आकार पर किसी भी सीएनएन के निष्पादन समय का अनुमान लगाने के लिए डिज़ाइन की गई एक पद्धति है। इन्फर प्रतिबंधित डीपीयू कॉन्फ़िगरेशन के भीतर काम करता है, केवल तीन डीपीयू को एक साथ सक्रिय करने की अनुमति देता है और ज़ाइलिंक्स टूल द्वारा बाहरी मेमोरी में प्रदान किए गए डिफ़ॉल्ट बस इंटरकनेक्शन को नियोजित करता है। इसके अतिरिक्त, यह कई डीपीयू के समवर्ती उपयोग के परिणामस्वरूप मेमोरी हस्तक्षेप के कारण आवश्यक अतिरिक्त समय का अनुमान लगा सकता है। डीपीयू पर चलने वाले सीएनएन के लिए ऊर्जा अनुमान इस थीसिस का एक और फोकस है। एनर्जीएनएन नामक एक ऊर्जा आकलन तकनीक सीएनएन और डीपीयू की विशेषताओं का लाभ उठाकर विकसित की गई है। इस तकनीक को मॉडल विकास में उपयोग नहीं किए गए नए सीएनएन की ऊर्जा खपत का अनुमान लगाने के लिए भी लागू किया जा सकता है। व्यापक मूल्यांकन इन दृष्टिकोणों की प्रभावशीलता को प्रदर्शित करता है, जिसमें निष्पादन समय के लिए छह दशमलव छह प्रतिशत की औसत भविष्यवाणी त्रुटियां और ऊर्जा अनुमान के लिए आठ दशमलव आठ प्रतिशत शामिल हैं। इन भविष्यवाणियों की उपयोगिता को यातायात निगरानी और ड्रोन प्रणालियों में वास्तविक दुनिया के अनुप्रयोगों के माध्यम से और भी स्पष्ट किया गया है।

थीसिस बड़ी संख्या में डीपीयू से युक्त डिज़ाइन स्थान का पता लगाने के लिए अपने दायरे का विस्तार करती है, इस प्रकार प्रभावी ढंग से समवर्तीता को बढ़ाती है। हालाँकि, यह विस्तार बस कनेक्शन और सीपीयू कोर सीमाओं से संबंधित जटिलताओं का परिचय देता है। इन चुनौतियों का समाधान करने के लिए, थीसिस विभिन्न कार्यों की प्रदर्शन मांगों को पूरा करने के लिए डीपीयू कॉन्फ़िगरेशन को अनुकूलित करने के लिए डिज़ाइन किए गए निष्पादन समय भविष्यवक्ता का परिचय देती है। समवर्ती सीएनएन को एक दूसरे और ऑपरेटिंग सिस्टम से अलग करने के लिए विभिन्न तरीकों को नियोजित किया जाता है। सीएनएन, डीपीयू और बस की विशेषताओं पर विचार करते हुए, डीपीयू कॉन्फ़िगरेशन पर दिए गए सीएनएन के निष्पादन समय की भविष्यवाणी करने के लिए एक्सप्रेस नामक एक मशीन लर्निंग-आधारित भविष्यवाणी दृष्टिकोण पेश किया जाता है। एक्सप्रेस सीपीयू और डीपीयू दोनों प्रोसेसिंग समय के लिए पूर्वानुमान प्रदान करता है, जिसके परिणामस्वरूप एंड-टू-एंड प्रोसेसिंग समय का अनुमान होता है।

इस दृष्टिकोण को और बढ़ाने के लिए, एक्सप्रेस-दो पेश किया गया है, जो विषम सीएनएन के लिए समर्थन प्रदान करता है। डीपीयू/सीएनएन की अलग-अलग संख्या के लिए अलग-अलग संख्या में सुविधाओं से बचने के लिए, एक्सप्रेस-दो सह-धावकों (सीएनएन जो सीएनएन के साथ समवर्ती रूप से चलते हैं जिसके लिए हम निष्पादन समय की भविष्यवाणी करते हैं) की सुविधाओं को एक साथ समेकित करता है। ऑपरेटिंग सिस्टम और वास्तविक सीएनएन को निष्पादित करने के लिए जिम्मेदार सीपीयू कोर को अलग करने के लिए एक नियंत्रक विकसित किया गया है। यह अलगाव निष्पादन समय माप में भिन्नता को कम करता है और भविष्यवाणी सटीकता को बढ़ाता है। इन सभी पद्धतियों में, भविष्यवाणी के लिए मशीन लर्निंग आधारित प्रतिगमन तकनीकों को नियोजित किया जाता है। सभी प्रयोग एक वास्तविक एफपीजीए बोर्ड पर आयोजित किए जाते हैं, और सोलह मानक सीएनएन की विशेषता वाले मूल्यांकन से बहुत कम भविष्यवाणी त्रुटियों का पता चलता है। एक्सप्रेस और एक्सप्रेस-दो क्रमशः दो दशमलव दो प्रतिशत और शून्य दशमलव सात प्रतिशत की औसत भविष्यवाणी त्रुटि प्राप्त करते हैं। फ्रेमवर्क की कम भविष्यवाणी त्रुटि इसे डिज़ाइन स्पेस अन्वेषण के लिए अत्यधिक प्रभावी बनाती है, जो एम्बेडेड सिस्टम एप्लिकेशन डेवलपर्स को महत्वपूर्ण लाभ प्रदान करती है।

Contents

List of Figures

List of Tables

List of Abbreviations

Sl. No.	Abbreviation	Expansion	Primarily defined in
1	ADAS	Advanced Driver Assistance System	Chap. 1
2	BRAM	Block Random Access Memory	Sec. 1.3
3	CNN	Convolutional Neural Network	Sec. 2.1.1
4	DPU	Deep learning Processor Unit	Sec. 2.1.2
5	DSE	Design Space Exploration	Sec. 1.3
6	DSP	Digital Signal Processor	Sec. 1.3
7	FFT	Fast Fourier Transform	Sec. 2.2
8	FPS	Frames Per Second	Sec. 1.3
9	FPGA	Field Programmable Gate Array	Chap. 1
10	GEMM	General Matrix Multiply	Sec. 2.2
11	HLS	High-level Synthesis	Sec. 2.2
12	MAVI	Mobility Assistant for Visually Impaired	Chap. 1
13	MGT	Multi-gigabit transceiver	Sec. 3.4
14	ML	Machine learning	Sec. 2.1.4
15	MPSoC	Multi-Processor System-on-Chip	Sec. 2.1.3
16	MSE	Mean Square Error	Sec. 2.1.4
17	PL	Programmable Logic	Sec. 2.1.3
18	PS	Processing System	Sec. 2.1.3
19	RNN	Recurrent Neural Network	Sec. 7.4
20	SVR	Support Vector Regression	Sec. 2.1.4
21	VHDL	VHSIC (Very High Speed Integrated Circuit) Hardware Description Language	Sec. 2.2