AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

DZIEDZINA: NAUKI INŻYNIERYJNO-TECHNICZNE
DYSCYPLINA: AUTOMATYKA, ELEKTRONIKA I ELEKTROTECHNIKA

ROZPRAWA DOKTORSKA

Sprzętowa akceleracja wymagających obliczeniowo operacji na potrzeby algorytmów sztucznej inteligencji w układach FPGA

Autor: mgr inż. Michał KARWATOWSKI

Promotor: Prof. dr hab. inż. Kazimierz WIATR
Promotor pomocniczy: dr hab. inż. Maciej WIELGOSZ

Praca wykonana:
Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie,
Wydział Informatyki, Elektroniki i Telekomunikacji, Instytut Elektroniki

Kraków, 2022

Pragnę złożyć podziękowania dla Pana
Profesora Kazimierza Wiatra oraz Pana
Doktora Macieja Wielgosza za nieocenioną
pomoc oraz wsparcie w trakcie prowadzenia
badań jak również mobilizację do pisania
niniejszej rozprawy.

Streszczenie

Niniejsza rozprawa dotyczy efektywnej implementacji algorytmów sztucznej inteligencji, w szczególności sieci neuronowych, w układach FPGA. Optymalna akceleracja sprzętowa pozwala poszerzyć zastosowania sztucznej inteligencji. Istotnym aspektem towarzyszącym badaniom prezentowanym w pracy jest optymalizacja energetyczna. Na akcelerację obliczeń składa się nie tylko dopasowana architektura sprzętowa, ale również optymalizacja algorytmów. W rozprawie badane są wpływy pruningu oraz kwantyzacji na algorytmy uczenia maszynowego, w tym sieci neuronowych. Głównymi obszarami zastosowań badanych algorytmów jest przetwarzanie języka naturalnego oraz analiza szeregów czasowych. Opracowane zostało narzędzie DL2HDL, które w istotny sposób ułatwia akcelerację w układach FPGA sieci neuronowych, zaprojektowanych w bibliotekach wysokopoziomowych, takich jak PyTorch. Konstrukcja oraz interface'y narzędzia zostały zaprojektowane w sposób przyjazny dla użytkownika, który nie posiada specjalistycznej wiedzy o projektowaniu akceleratorów FPGA. Ważnym elementem narzędzia DL2HDL jest możliwość symulacji oraz testowania działania algorytmu na każdym kroku implementacji projektu. Dzięki opracowanym optymalizacjom narzędzie generuje wydajny akcelerator zorientowany na niską latencję. Optymalizacje pod kątem latencji są wystarczająco daleko idące, aby umożliwić zastosowanie sieci neuronowych w systemach wspomagających akcelerator cząstek LHC pracujący w CERN. Cel rozprawy został osiągnięty. Badania pozwoliły na określenie optymalnych poziomów pruningu oraz kwantyzacji dla analizowanych algorytmów. Wypracowana została lista wymagań oraz powstała implementacja narzędzia DL2HDL mapującego wysokopoziomowy opis sieci neuronowych do zoptymalizowanych architektur sprzętowych w układach FPGA. Szeroko zakrojone optymalizacje pozwoliły na osiągnięcia najniższej spotkanej dotąd w literaturze latencji w trakcie inferencji rekurencyjnych sieci neuronowych typu LSTM.

Abstract

This dissertation concerns the effective implementation of artificial intelligence algorithms, in particular neural networks, in FPGAs. Optimal hardware acceleration allows artificial intelligence algorithms to expand the field of applications. An important aspect accompanying the research presented in this dissertation is energy optimization. The acceleration of calculations consists not only of the adapted hardware architecture but also the optimization of algorithms. The dissertation examines the effects of pruning and quantization on machine learning algorithms, including neural networks. The main areas of application of the analyzed algorithms are natural language processing and time series analysis. The DL2HDL tool has been developed, which significantly facilitates the acceleration of neural networks designed in high-level libraries, such as PyTorch. The design and interfaces of the tool have been designed in the most user-friendly way, and does not require from the user any specialist knowledge about designing FPGA accelerators. An important element of the tool is the ability to simulate and test the operation of the algorithm on each step of the project implementation. Thanks to the developed optimizations, the tool generates an efficient accelerator focused on low latency. Latency optimizations go far enough to enable the use of neural networks in systems supporting the LHC particle accelerator at the CERN research center. The goal of the dissertation was successfully achieved. The research allowed to determine the optimal levels of pruning and quantization for the analyzed algorithms. A list of requirements was compiled and an implementation of a tool mapping a high-level description of neural networks to optimized hardware architectures in FPGAs was developed. Aggressive optimizations allowed to achieve state-of-the-art latency during the inference of LSTM recurrent neural networks.

Spis treści