Częstochowa, dn. 31.05.2021

Prof. dr hab. inż. Roman Wyrzykowski
Katedra Informatyki
Politechnika Częstochowska
ul. Dąbrowskiego 69
42-201 Częstochowa
roman@icis.pcz.pl

RECENZJA
ROZPRAWY DOKTORSKIEJ
mgr inż. Dominika Żurka
"Alceleracja obliczeń algorytmów uczenia maszynowego oraz wybranych populacyjnych algorytmów inteligencji obliczeniowej ze zredukowną precyzją danych poprzez implementację w układach GPGPU"

Promotor: Prof. dr hab. inż. Kazimierz Wiatr
Katedra Informatyki
Wydział Informatyki, Elektroniki i Telekomunikacji
Akademia Górniczo-Hutnicza im. St. Staszica w Krakowie

1. Problem badawczy i jego znaczenie

Tematyka przedłożonej do recenzji rozprawy doktorskiej mgr inż. Dominika Żurka dotyczy bardzo ważnego i aktualnego obszaru badawczego łączącego obliczenia równoległe, które w kontekście pracy można utożsamiać z technologią HPC, oraz inteligencję obliczeniowej (ang. computational intelligence), przy czy ta ostatnia ściśle powiązana jest z niesłychanie dzisiaj nośną dziedziną sztucznej inteligencji (ang. artificial intelligence - AI). Chociaż każda z tych dziedzin, tj. HPC oraz AI, do pewnego momentu rozwijała się niezależnie, to w okresie ostatnich około dziesięciu lat zaczęły się one splatać w sposób wręcz nierozerwalny, stymulując się wzajemnie. Pomimo wielu spektakularnych osiągnięć, które stworzyły m.in. podwaliny tzw. czwartej rewolucji przemysłowej, dalszy rozwój w rozpatrywanym obszarze napotyka na szereg barier, których pokonanie wymaga kompleksowych działań zarówno w sferze badań podstawowych, jak i stosowanych, a także technologii mikroelektronicznej.

Z racji mojej specjalizacji naukowej bliższe są mi bariery w dziedzinie HPC, wynikające chociażby z ograniczeń technologicznych w mikroelektronice, prowadzące do powstania barier związanych z szybkością dostępu do danych, wydajnością

komunikacji czy też zapotrzebowania na energię zużywaną przez systemy komputerowe. Chęć przełamania tych barier doprowadziła już w naszym wieku do wprowadzenia i upowszechnienia architektur wielo- i masywnie wielordzeniowych. Ważnym krokiem w tym kierunku stało się również coraz szersze wykorzystanie systemów obliczeniowych o architekturze heterogenicznej, w skład których wchodzi wielopoziomowa struktura łącząca urządzenia o różnych charakterystykach, np. procesory CPU ogólnego przeznaczenia, różnego typu akceleratory obliczeniowe czy też układy programowalne FPGA. W ostatnim okresie coraz bardziej popularne stają się też architektury specjalizowane, dedykowane do określonej dziedziny zastosowań (ang. Domain Specific Accelerators – DSA). Jest to szczególnie widoczne w obszarze uczenia maszynowego i sztucznej inteligencji, gdzie swoistym przełomem było opracowanie przez firmę Google pierwszej wersji układu TPU (Tensor Processing Unit), dedykowanego do realizacji fazy inferencji funkcjonowania sieci neuronowych, a następnie kolejnych układów TPU zorientowanych na implementację jeszcze bardziej wymagającej obliczeniowo fazy uczenia głębokich sieci neuronowych.

Recenzowana praca mgra inż. Dominika Żurka dotyczy akceleratorów graficznych, znanych również jako procesory GPU i poświęcona jest dostatecznie wszechstronnemu zbadaniu perspektyw efektywnej implementacji wybranych algorytmów inteligencji obliczeniowej i sztucznej inteligencji z zastosowaniem tych akceleratorów, ze szczególnym uwzględnieniem możliwości, jakie w tym zakresie daje wykorzystanie zredukowanej precyzji danych. Możliwości te obejmują nie tylko zwiększenie szybkości przetwarzania, na czym skoncentrował się doktorant, lecz również zmniejszenie zużycia energii niezbędnej do wykonania obliczeń. Ten drugi efekt przekłada się nie tylko na mniejsze rachunki za energię elektryczną, lecz również na coraz ważniejsze w skali całej ludzkości zmniejszenie negatywnego efektu oddziaływania na otaczające nas środowisko, szczególnie w postaci tzw. śladu węglowego (ang. carbon footprint). Zaproponowane w pracy synergiczne podejście do problemu oparte na uwzględnieniu właściwości zarówno samych algorytmów i metod inteligencji obliczeniowej oraz architektur systemów obliczeniowych korzystnie świadczy o przygotowaniu autora do działalności badawczej i jego zdolności do rozwiazywania dostatecznie złożonych zagadnień badawczych.

Podsumowując powyższe uwagi, kierunek badań, wybór problematyki rozprawy oraz jej tezy i celów zaproponowane przez mgra inż. Dominika Żurka oceniam zdecydowanie pozytywnie. Lokują się one korzystnie w nakreślonej wyżej tematyce współczesnej informatyki, dotyczącej podstaw teoretycznych i zastosowań praktycznych obliczeń równoległych i akceleratorów w dynamicznie rozwijającym się obszarze inteligencji obliczeniowej czy też sztucznej inteligencji, definiując zarówno teoretyczny, jak i przede wszystkim praktyczny charakter rozprawy. Według mojej najlepszej wiedzy, rozprawa ta jest faktycznie pierwszą próbą w obszarze krajowym

zmierzenia się z daną tematyką w kontekście wykorzystania zredukowanej precyzji danych i akceleratorów GPU.

2. Koncepcja i redakcja rozprawy

Recenzowana praca doktorska obejmuje formalnie 7 rozdziałów oraz stosowną bibliografię zawierającą 129 pozycji. Rozprawa liczy łącznie 137 stron. Dodatkowo jako dodatek do rozprawy została załączona płyta DVD zawierająca implementacje opracowane w ramach pracy.

W ***rozdziale pierwszym*** Autor zawarł motywację i cele badań, a także sformułował tezę pracy. Chociaż pierwsza część tezy dotycząca możliwości przyspieszenia obliczeń dzięki wykorzystaniu akceleratorów graficznych jest mało oryginalna na tle dotychczasowego stanu badań prowadzonych w rozpatrywanej dziedzinie od wielu lat, znacznie bardziej interesująca jest druga część tezy odnosząca się do możliwości wykorzystania zredukowanej precyzji danych w obliczeniach na kartach graficznych do przyspieszenia obliczeń, w tym również bez pogorszenia jakości algorytmu. Rozdział kończy się przedstawieniem treści pracy zawartej w kolejnych rozdziałach.

Rozdział drugi poświecono dostatecznie zwięzłemu wprowadzeniu do obszernych zagadnień sztucznej inteligencji, inteligencji obliczeniowej oraz uczenia maszynowego. Szczególną uwagę poświęcono przy tym dziedzinom rozpatrywanym w niniejszej rozprawie, a mianowicie ewolucyjnym algorytmom agentowym i sieciom neuronowym, ze szczególnym uwzględnieniem bardzo popularnych w zastosowaniach praktycznych sieci konwolucyjnych oraz rekurencyjnych. Rozdział ten wydaje się też najbardziej odpowiednim miejscem do wprowadzenia algorytmu wektorów nośnych (ang. Support Vector Machine -SVM) – ważnego algorytmu uczenia maszynowego, którego implementacja stanowi przedmiot rozdziału czwartego. Niestety Autor odłożył charakterystykę tego algorytmu do tego właśnie rozdziału, co wydaje się mniej logicznym rozwiązaniem.

Kolejny, trzeci rozdział rozprawy zawiera charakterystykę platform sprzętowych wykorzystywanych w pracy. W praktyce skoncentrowano się na przedstawieniu architektury i właściwości układów graficznych, a w szczególności kart graficznych NVIDIA Volta oraz platformy NVIDIA CUDA oferującej m.in. środowisko programistyczne dla układów GPU. Rozdział ten traktuje tę obszerną problematykę bardzo pobieżnie. W szczególności, pewne zdziwienie budzi brak właśnie w tym rozdziale szerszego omówienia kluczowego dla tematyki pracy zagadnienia wsparcia sprzętowego w procesorach GPU formatów numerycznych wprowadzonych w celu efektywnej realizacji obliczeń ze zredukowaną precyzją.

Rozdział czwarty rozpoczyna w pełni oryginalną część dysertacji, prezentując zaproponowane przez doktoranta rozwiązanie dla implementacji w układach GPU zagadnienia wyznaczania sekwencji binarnych z minimalną wartością funkcji autokorelacji (ang. low autocorrelation binary sequence – LABS) z wykorzystaniem algorytmów heurystycznych w tym algorytmów SDLS oraz *tabu search*. Algorytmy te wiążą się z realizacją wspomnianych wcześniej zagadnień ewolucyjnych algorytmów agentowych wykorzystywanych do rozwiązywania problemów optymalizacji lokalnej. Wyrazić tutaj należy pewien żal, iż np. w rozdziale drugim, gdzie wprowadzono algorytmy agentowe, autor nie znalazł miejsca na chociażby pobieżne scharakteryzowanie ich związku z zagadnieniem wyznaczania wspomnianej funkcji autokorelacji, co utrudnia śledzenie treści rozprawy i ocenę podjętych badań. W szczególności, dotyczy to oceny racjonalności i wręcz konieczności koncentrowania się na wyznaczaniu tej funkcji.

Rozdział piąty stanowi moim zdaniem najbardziej wartościowy fragment rozprawy. Poświęcono go zagadnieniom przetwarzania języka naturalnego, a dokładniej – wykorzystaniu algorytmu wektorów nośnych (ang. Support Vector Machines) do klasyfikacji tekstów. Właśnie ten rozdział najbardziej spełnia moje oczekiwania związane z tytułem pracy, związane z możliwością zastosowania zredukowanej precyzji danych (i tym samym obliczeń) do zwiększenia efektywności realizacji algorytmów inteligencji obliczeniowej. Jak stwierdza sam doktorant, podstawowym celem tego rozdziału jest wykazanie wpływu dwóch zaproponowanych metod kwantyzacji wykorzystywanych w procesie trenowania wektorów nośnych na efektywność algorytmu klasyfikacji. W szczególności, umożliwiło to zbadanie wpływu zredukowanej precyzji danych na szybkość realizacji algorytmu klasyfikacji w układach GPU.

W ***rozdziale szóstym*** doktorant kontynuuje podjęty w poprzednim rozdziale wątek wpływu zredukowanej precyzji obliczeń na szybkość przetwarzania algorytmu w procesorach GPU, koncentrując się na obliczaniu operacji konwolucji (znanych też pod nazwą operacji splotowych), szeroko wykorzystywanych w realizacji algorytmów głębokiego uczenia w oparciu o konwolucyjne sieci neuronowe (ang. convolutional neural networks – CNN). W tym przypadku dodatkowo zastosowano techniki redukcji objętości sieci (ang. pruning) oraz zaproponowany algorytm obliczania konwolucji nazwany *konwolucją rzadką*. Celem podjętych badań jest opracowanie implementacji operacji splotowych w sposób efektywniejszy niż w standardowej bibliotece cuDNN udostępnianej przez firmę NVIDIA.

W ***rozdziale siódmym*** dokonano zwięzłego podsumowania dysertacji, w tym jej oryginalnych elementów, a także wskazano na obiecujące perspektywy rozwoju uzyskanych rezultatów w ramach dalszych prac Autora. W szczególności, mogą one

stanowić zalążek nowej biblioteki umożliwiającej efektywne przeprowadzanie obliczeń z dziedziny inteligencji obliczeniowej (sztucznej inteligencji).

3. Wkład autora

Uwzględniając powyższe omówienie zawartości pracy oraz ogólną pozytywną ocenę jej zawartości merytorycznej, do najważniejszych wyników pracy i osiągnięć autora należy zaliczyć:

1. Podstawowym wynikiem o ogólnym charakterze jest wykazanie możliwości istotnego zwiększenia efektywności obliczeń równoległych w dziedzinie inteligencji obliczeniowej i uczenia maszynowego realizowanych z zastosowaniem akceleratorów GPU dzięki wykorzystaniu zredukowanej precyzji danych. Uzyskane korzyści obejmują nie tylko przyspieszenie obliczeń, lecz również oszczędność wymaganej pamięci, co umożliwia wydajne rozwiazywanie zagadnień o większych rozmiarach.
2. Osiągniecie powyższego celu w największym stopniu przejawiło się w wykazanie wpływu dwóch zaproponowanych metod kwantyzacji wykorzystywanych w procesie trenowania wektorów nośnych algorytmu SVM na efektywność algorytmu klasyfikacji tekstu. Ważną cechą zaproponowanych metod jest możliwość elastycznego sterowania jakością klasyfikatora SVM.
3. Kolejnym istotnym innowacyjnym elementem pracy, niezbędnym do uzyskania wyniku wspomnianego w punkcie 1, jest także wszechstronna analiza możliwości przyspieszenia wykonania na kartach graficznych operacji konwolucji, które są szeroko wykorzystywane w realizacji algorytmów głębokiego uczenia opartych o konwolucyjne sieci neuronowe, dzięki zastosowaniu techniki pruningu i operacji na macierzach rzadkich. Analiza ta została poparta opracowaniem autorskiej implemantacji na GPU pozwalającej na uzyskanie wydajności lepszej niż oferowana przez standardową bibliotekę cuDNN.
4. W zakresie akceleracji obliczeń w ewolucyjnych algorytmach agentowych inteligencji obliczeniowej wykorzystywanych do rozwiązywania problemów optymalizacji lokalnej na uwagę zasługują także zaproponowane oryginalne algorytmy wyznaczania funkcji autokorelacji sekwencji binarnych, dzięki czemu autorowi udało się zapewnić istotne przyspieszenie obliczeń na CPU w stosunku do procesorów ogólnego przeznaczenia.

Zgodnie z bazą SCOPUS, uzyskane wyniki zostały opublikowane w 8 pracach w języku angielskim, przy czym jedna publikacja ukazała się w jednym bardzo dobrym czasopiśmie indeksowanym w bazie JCR (Journal of Computational Science – 100 punktów Ministerstwa, współczynnik wpływu IF = 2.6), dwie – w specjalistycznych

czasopismach międzynarodowych, zaś cztery prace z udziałem autora opublikowano w materiałach konferencji międzynarodowych. Świadczy to bardzo pozytywnie o stopniu weryfikacji uzyskanych rezultatów przez międzynarodową społeczność specjalistów zajmujących się rozpatrywaną dziedziną.

4. Poprawność pracy i uwagi krytyczne

Poprawność treści pracy nie wzbudza moich istotnych zastrzeżeń, a stwierdzenia w niej zawarte wydają się być godne zaufania, co wynika w szczególności z dosyć szczegółowych uzasadnień, potwierdzonych wynikami przeprowadzonych badań eksperymentalnych, do których wykorzystano w szczególności procesory graficzne firmy NVIDIA typu GeForce GTX 1060 (architektura Pascal) oraz Tesla V100 (architektura Volta). Merytorycznie i metodologicznie jest ona generalnie poprawna, pozostaje jednak szereg kwestii, które wymagają wyjaśnienia:

1. Pierwszą część tezy rozprawy dotycząca możliwości przyspieszenia obliczeń dzięki wykorzystaniu akceleratorów graficznych należy uznać za mało oryginalna na tle dotychczasowego stanu badań prowadzonych w rozpatrywanej dziedzinie od wielu już lat. Jednakże powyższy wniosek związany z przewagą specjalizowanych procesorów GPU w stosunku do procesorów ogólnego przeznaczenia (CPU) nie jest już tak oczywisty w kontekście innego ważnego parametru charakteryzującego obliczenia, którym jest zużycie energii w ich trakcie. W związku z powyższym nasuwa się pytanie o porównanie GPU i CPU w rozpatrywanych w pracy zastosowaniach z uwzględnieniem obydwu parametrów – wydajności oraz kosztów energetycznych.
2. Jak już wspomniano przy omówieniu treści pracy, w rozdziale trzecim rozprawy zabrakło szerszego omówienia kluczowego dla jej tematyki zagadnienia wsparcia sprzętowego w procesorach GPU formatów numerycznych wprowadzonych w tych procesorach w celu efektywnej realizacji obliczeń ze zredukowaną precyzją. Generalnie rozdział poświęcony obliczeniom równoległym i układom GPU wydaje się zbyt pobieżnie traktować tę problematykę w relacji do jej wagi w kontekście celu rozprawy. Brakuje mi w nim również odwołania do podstawowej dla omawianych problemów klasycznej monografii autorów Hennesy i Pattersona pt. „Computer Architecture: A Quantitative Approach", którego najnowsze wydanie ukazało się na przełomie lat 2017 i 2018. W monografii tej wypunktowano m.in. ścisły związek architektur GPU z komputerami wektorowymi, co pozwala lepiej zrozumieć właściwości tych architektur i kierunki ich rozwoju.
3. Wśród innych uwag, które zawarłem w omówieniu treści rozprawy i które utrudniają śledzenie treści rozprawy i ocenę podjętych badań, chciałem w pierwszej kolejności zwrócić uwagę na brak odpowiedniej charakterystyki miejsca i roli funkcji autokorelacji sekwencji binarnych w kontekście algorytmów agentowych, co jest wręcz niezbędne dla uzasadnienia racjonalności i wręcz konieczności koncentrowania się na wyznaczaniu tej właśnie funkcji. Po

drugie, w rozdziale piątym przy rozpatrywaniu zagadnienia klasyfikacji tekstu arbitralnie, bez żadnej wstępnej analizy np. literaturowej, przyjęto wykorzystanie w tym celu podejścia opartego o zastosowanie algorytmu wektorów nośnych, podczas gdy w wielu pracach bardzo dobre wyniki w zakresie przetwarzania języka naturalnego (ang. natural language processing) osiągnięto dzięki zastosowaniu metod opartych na uczeniu głębokim i sieciach neuronowych.

4. Istotną słabością rozdziału szóstego rozprawy, w którym autor koncentruje się na poprawie efektywności operacji konwolucji w układach GPU, jest brak porównania uzyskanych wyników z rezultatami prac innych autorów. Autor porównuje uzyskane przez siebie wyniki tylko z biblioteką cuDNN firmy NVIDIA. Tymczasem w okresie ostatnich kilku lat problematyka ta stała się przedmiotem intensywnych badań prowadzonych przez wielu autorów w różnych krajach, co pozwoliło na osiągnięcie dużego postępu w stosunku do biblioteki cuDNN. Świadczą o tym m.in. obszerne przeglądy publikowane przez specjalizującego się w tego typu publikacjach indyjskiego autora Sparsh Mittala. Rozumiejąc trudność z odtworzeniem wyników tych prac w celu przeprowadzenia dogłębnej analizy porównawczej, niezrozumiały jest jednak brak odniesienia się w rozprawie do metod proponowanych w pracach tych autorów w celu obiektywnego wypunktowania oryginalności uzyskanych wyników w stosunku do istniejącego stanu wiedzy. Dotyczy to chociażby zastosowania techniki pruningu czy też macierzy rzadkich, które stosowane są przez wielu badaczy.

5. W przypadku dowolnych badań eksperymentalnych istotne znaczenie dla zapewnienia wiarygodności otrzymanych wyników posiada przeprowadzenie właściwej analizy statystycznej tych wyników. Jedyna informacja, jaką udało mi się znaleźć w rozprawie na ten temat, to fakt powtórzenia pomiarów czasu wykonania 10 razy. Nie wydaje się jednak, aby była to informacja wystarczająca, gdyż np. nie mówi ona nic o rozrzucie otrzymanych wyników pomiarów.

5. Konkluzja

Pomimo przedstawionych uwag krytycznych, do których doktorant powinien się odnieść na obronie pracy, rozprawę oceniam pozytywnie jako stanowiącą ciekawe rozwiązanie naukowego problemu zwiększenia efektywności implementacji algorytmów inteligencji obliczeniowej z wykorzystaniem współczesnych architektur równoległych. W mojej opinii rozprawa spełnia wymagania stawiane pracom doktorskim, zatem wnoszę o dopuszczenie mgr inż. Dominika Żurka do dalszych etapów przewodu doktorskiego.

Wyrzykowski R.