Poznań, 30-11-2022

dr hab. inż. Paweł Śniatała, prof. PP
Wydział Informatyki i Telekomunikacji
Politechniki Poznańskiej
ul. Piotrowo 3A, 60-965 Poznań
pawel.sniatala@put.poznan.pl

RECENZJA ROZPRAWY DOKTORSKIEJ

mgr inż. Michała Karwatowskiego

„SPRZĘTOWA AKCELERACJA WYMAGAJĄCYCH OBLICZENIOWO OPERACJI NA POTRZEBY ALGORYTMÓW SZTUCZNEJ INTELIGENCJI W UKŁADACH FPGA"

wykonana na podstawie umowy zawartej z Akademią Górniczo-Hutniczą im. Stanisława Staszica w Krakowie, Wydział Elektrotechniki, Automatyki, Informatyki i Inżynierii Biomedycznej.

1. *Charakterystyka wyboru tematu i przedmiot rozprawy*

Przedmiotem rozprawy doktorskiej mgr inż. Michała Karwatowskiego są zagadnienia związane z akceleracją sprzętową wybranych operacji wykorzystywanych w algorytmach sztucznej inteligencji (AI). Algorytmy tej klasy wymagają dużej mocy obliczeniowej i często przekraczają granice możliwości komercyjnie dostępnego sprzętu. Stąd też, w pierwszym etapie naturalnym kierunkiem było wykorzystanie mikroprocesorów wielordzeniowych lub kart graficznych (GPU). Karty GPU zawierają tysiące małych jednostek obliczeniowych działających równolegle i wykonują wiele niskopoziomowych operacji matematycznych dla takich efektów renderingu jak cienie, odbicia, oświetlenie i przezroczystość. Możliwość równoległego wykonywania obliczeń matematycznych świetnie pasuje do operacji wykonywanych w ramach algorytmów AI np. algorytmów głębokiego uczenia. Kolejnym krokiem rozwoju technik obliczeniowych związanych z AI było opracowanie architektur sprzętowych dedykowanych do specyfiki obliczeń wykonywanych w algorytmach AI. Należą do nich jednostki przetwarzające sieci neuronowe (Neural Network Processing Units - NNPU), jednostki przetwarzające tensory (Tensor Processing Units - TPU), układy

neuromorficzne (oxide-based memristors, spintronic memories, threshold switches, transistors) i wreszcie układy programowalne FPGA, które są wykorzystywane w rozwiązaniach prezentowanych w recenzowanej rozprawie.

W ciągu ostatnich kilku lat można zauważyć imponująco szybki rozwój architektur sprzętowych zoptymalizowanych pod kątem uczenia maszynowego, uczenia głębokiego, wizji komputerowej, przetwarzania języka naturalnego i innych zadań związanych z AI. Oczekuje się, że w najbliższej przyszłości globalny rynek układów scalonych dla sztucznej inteligencji osiągnie bardzo wysoką trajektorię wzrostu. Prace badawcze doktoranta idealnie wpisują się w ten trend.

Specyfika obliczeń związanych z algorytmami sztucznej inteligencji wymusza określone wymogi projektowanego sprzętu. Głębokie sieci neuronowe mają więcej operacji mnożenia macierzy i konwolucji. Ponieważ operacje te różnią się od tradycyjnych obciążeń procesora, wielopoziomowe pamięci podręczne nie są potrzebne, co skutkuje prostszą pamięcią na chipie. W przypadku zadań związanych z percepcją sensoryczną, takich jak widzenie komputerowe, tłumaczenie języków i rozpoznawanie mowy, rzadko wymagana jest pełna precyzja. Większość wnioskowania w czasie rzeczywistym może być wykonywana przy użyciu wielkości 8- lub 16-bitowych. Zmniejszając precyzję numeryczną można zwiększyć przepustowość obliczeniową. Najnowocześniejszy sprzęt AI odchodzi od tradycyjnych 32-bitowych obliczeń o pojedynczej precyzji na rzecz 8- lub 16-bitowej precyzji, oszczędzając w ten sposób powierzchnię układu scalonego i umożliwiając wykorzystanie większej liczby jednostek arytmetycznych w celu maksymalizacji przetwarzania równoległego. Istotnym elementem jest również optymalne zużycie energii i wydajność. Obciążenia obliczeń algorytmów AI obejmują zarówno etap trenowania/uczenia, który wymaga dużej pamięci masowej, jak i intensywne obliczeniowo wnioskowanie. Dużym wyzwaniem jest więc osiągnięcie wydajnego uczenia i wnioskowania na tym samym urządzeniu przy optymalnym zużyciu energii i wydajności.

Mierniki wydajności obliczeniowej akceleratorów sprzętowych (liczba operacji na sekundę np. TOPS - Tera Operations Per Second) oraz wydajności energetycznej (TOPS/Wat) są więc kluczowymi parametrami proponowanych rozwiązań. Kolejną ważną cechą akceleratorów sprzętowych jest łatwość programowania, co przekłada się na produktywność inżynierów korzystających z tych rozwiązań. Programowalność umożliwia szybkie tworzenie nowych konfiguracji konwolucyjnych sieci neuronowych (CNN), sieci neuronowych rekurencyjnych (RNN), sieci z pamięcią długotrwałą (LSTM) oraz sieci uczących się przez wzmocnienie. Większość uczenia maszynowego jest zwykle implementowana w centrach danych w

chmurze. W przypadku niektórych zastosowań bardziej sensowne jest zastosowanie uczenia maszynowego na brzegu sieci (edge ecosystem), niż podłączenie z powrotem do chmury. Główną korzyścią jest mniejsze zużycie pasma i minimalizacja opóźnień. Architektury akceleratorów sprzętowych AI muszą więc być dobrze dopasowane do konkretnych ról w ekosystemie. W rozprawie, doktorant podejmuje w szczególności zagadnienia związane z wydajnością obliczeniową i energetyczną proponowanych rozwiązań, uwzględniając wspomniane powyżej mierniki jakości tych implementacji. Podsumowując tą część recenzji, należy wysoko ocenić aktualność poruszanych w rozprawie zagadnień i trafny wybór obszaru badań, który wpisuje się w zakres bieżąco dyskutowanych problemów badawczych. Doktorant jasno sformułował tezy, które następnie w sposób klarowny udowodnił poprzez analizę przeprowadzoną w rozprawie.

Tezy rozprawy, w formie przedstawionej przez Autora są następujące:

1. „Delegacja części obliczeń algorytmów sztucznej inteligencji do akceleratorów sprzętowych w układach FPGA pozwoli na optymalizację energetyczną obliczeń oraz skrócenie czasu wykonywania algorytmów, co pozwoli na poszerzenie pola zastosowań algorytmów sztucznej inteligencji."
2. „Możliwa jest agresywna optymalizacja inferencji algorytmów uczenia maszynowego oraz efektywna ich implementacja w układach FPGA pozwalająca na osiąganie ekstremalnie niskich latencji przy jednoczesnym zachowaniu wysokiej jakości wyników predykcji."
3. „Zaprojektowanie narzędzia pozwalającego na konwersję wysokopoziomowego opisu sieci neuronowych do silnie zoptymalizowanych implementacji w układach FPGA pozwoli na łatwą delegację obliczeń do akceleratorów sprzętowych."

2. *Ocena merytoryczna rozprawy*

Praca składa się z sześciu rozdziałów, spisu literatury oraz dwóch załączników. Pierwszy załącznik stanowi opis procedury transferu wytrenowanej sieci neuronowej do innego, wybranego przez użytkownika, narzędzia komputerowego. Drugi dodatek zawiera przykładowy kod, który realizuje symulację i konwersję rekurencyjnej sieci neuronowej. Spis literatury zawiera 82 pozycje wraz z publikacjami Autora, powiązanymi z przedstawianą w rozprawie tematyką. Rozprawa posiada uporządkowaną logicznie strukturę, która jasno przedstawia analizę prowadzącą do wykazania słuszności sformułowanych tez pracy.

Rozdziały pierwszy i drugi stanowią teoretyczne wprowadzenie w tematykę rozprawy. W tym miejscu, między innymi, przedstawione są zagadnienia związane z redukcją precyzji

obliczeń, czyli kwantyzacja i pruning, które są implementowane w proponowanych w dalszej części pracy rozwiązaniach. Autor prezentuje również metryki, które posłużą do oceny proponowanych rozwiązań.

W kolejnym, trzecim rozdziale, przedstawione jest autorskie rozwiązanie akceleratora sprzętowego, który został wykorzystany do znajdowania dokumentów podobnych do referencyjnego. Zaprojektowane i przeprowadzone eksperymenty posłużyły autorowi do analizy porównawczej wydajności obliczeniowej oraz zużycia energii w proponowanych strukturach sprzętowych. Do eksperymentów użyto dwóch platform sprzętowych. Pierwsza to platforma ZedBoard, oparta na urządzeniu rodziny Xilinx Zynq-7000, która łączy wbudowany procesor ARM i FPGA. Platforma ta charakteryzuje się niskim poborem mocy. Drugą platformą sprzętową jest typowy węzeł klasy serwerowej, wykorzystujący czterordzeniowy procesor Intel Core-i7 950. Dodatkowo na serwerze znajdował się zestaw Xilinx Virtex-7 z matrycą FPGA VC7070. Uzyskane wyniki wykazały, że zastosowanie FPGA do przyspieszenia obliczeń znacząco podnosi zarówno wydajność (przyspieszenie rzędu 10.5 do 11.7 w stosunku do implementacji bez akceleratora), jak i efektywność energetyczną (zużycie energii zostało obniżone o 10.8 oraz 12.9 odpowiednio dla platformy o niskim i wyższy poborze mocy).

Kolejnym istotnym problemem naukowym, który został rozwiązany przez Autora rozprawy i przedstawiony w rozdziale czwartym, jest zaproponowana redukcja precyzji obliczeń oraz zbadanie wpływu ilości danych na korelację wyników metryki podobieństwa kosinusowego. Proponowane rozwiązania zostały zweryfikowane w oparciu o akcelerator sprzętowy, którego synteza do układu FPGA została wykonana przy użyciu syntezy wysokopoziomowej. Przeprowadzone eksperymenty pokazały, że podwójny format danych zmiennoprzecinkowych IEEE 754 może okazać się niekonieczny w niektórych zastosowaniach. Jak konkluduje Autor, dokładność można zmniejszyć z 8 bajtów do 1 bajta, zachowując 99% korelacji wyników. Dodatkowo można zmniejszyć rozmiar przetwarzanych danych, co ma duży wpływ na wymagania dotyczące pamięci i przepustowość transmisji danych.

W drugiej części rozdziału czwartego doktorant zweryfikował jak modyfikacja reprezentacji danych wpływa na dokładność algorytmu uczenia maszynowego struktur KNN (K najbliższych sąsiadów). W tym wypadki eksperymenty przeprowadzono na bazie odręcznie napisanych cyfr (MNIST). Do implementacji wykorzystano wysokopoziomowe narzędzia Vivado HLS. Przeprowadzone eksperymenty pokazują, że reprezentacja i wymiar danych mogą być nadmiarowe. Przeprowadzona redukcja znacząco obniżyła wykorzystanie zasobów.

Trzecia część rozdziału poświęcona jest technikom kompresji sieci neuronowych zaprojektowanych do działania z przetwarzaniem języka naturalnego. Autor stosuje pruning

oraz kwantyzację. W wyniku tych eksperymentów pokazano, że wielkość modelu można zmniejszyć o ponad 84% przy niewielkiej degradacji wydajności wynoszącej około 1%.

Obszerny pięćdziesięcioczterostronicowy rozdział piąty przedstawia prace badawcze autora dotyczące mapowania sieci neuronowych do języków opisu sprzętu (HDL). Autor opracował, zaimplementował i przetestował narzędzie DL2HDL, które pozwala na łatwą konwersję wysokopoziomowych opisów modeli neuronowych do wysoce zoptymalizowanego kodu HDL. Narzędzie przygotowane jest do optymalizacji zarówno architektur neuronowych opartych na warstwach liniowych jak i rekurencyjnych typu LSTM (Long Short-Term Memory). W procesie optymalizacji uwzględnione są najpopularniejsze funkcje aktywacyjne.

Struktura pracy jest jasna i klarowna, a dodając drobne uwagi krytyczne, zabrakło w treści rozprawy przedstawienia otrzymanych w wyniku syntezy bloków sprzętowych. Schematy blokowe pokazujące funkcjonalności poszczególnych części realizacji sprzętowej przybliżyłyby czytelnikom strukturę otrzymanego sprzętu.

Drobne błędy edycyjne zostały zebrane w osobnym pliku i przekazane doktorantowi do ewentualnego wykorzystania.

3. Wnioski końcowe

Pan mgr inż. Michał Karwatowski w recenzowanej rozprawie doktorskiej przedstawił rozwiązania kilku znaczących problemów badawczych związanych z optymalizacją akceleratorów sprzętowych wykorzystywanych w obliczeniach dedykowanych algorytmom sztucznej inteligencji. Niewątpliwe przedstawione rezultaty pracy Doktoranta stanowią oryginalne i wartościowe wyniki, które stanowią istotny wkład w przedmiotowej dziedzinie.

Istotnym walorem pracy jest również jej aspekt praktyczny. Opracowane narzędzie komputerowe może być wykorzystywane w pracach inżynierów i znaczenie zwiększyć wydajność końcowych rezultatów.

Należy również wspomnieć o znaczących publikacjach doktoranta. Wyszukiwanie w bazie bibliotecznej wykazuje 13 pozycji, które zostały opublikowane w czasopismach polskich i zagranicznych. Mając na uwadze przedstawione wyniki, zarówno teoretyczne jak i praktyczne, oraz biorącą pod uwagę dotychczasowy dorobek publikacyjny doktoranta, wnioskuję o wyróżnienie pracy doktorskiej Pana mgr inż. Michała Karwatowskiego.

Podsumowując niniejszą recenzję stwierdzam, że praca pt.: „SPRZĘTOWA AKCELERACJA WYMAGAJĄCYCH OBLICZENIOWO OPERACJI NA POTRZEBY ALGORYTMÓW SZTUCZNEJ INTELIGENCJI W UKŁADACH FPGA" spełnia wymagania stawiane rozprawom doktorskim w dyscyplinie elektronika, odpowiadającej dziedzinie nauk inżynieryjno-technicznych,

dyscyplinie Automatyka, Elektronika i Elektrotechnika wg klasyfikacji określonej w Rozporządzeniu MNiSzW z dnia 20 września 2018 roku w sprawie dziedzin nauki i dyscyplin naukowych oraz dyscyplin artystycznych (Dz.U.2018 poz1818) i wnoszę o dopuszczenie mgr inż. mgr inż. Michała Karwatowskiego do dalszych etapów przewodu doktorskiego.

Paweł Anders

Uwagi edycyjne dotyczące rozprawy doktorskiej, mgr inż. Michała Karwatowskiego

„SPRZĘTOWA AKCELERACJA WYMAGAJĄCYCH OBLICZENIOWO OPERACJI NA POTRZEBY ALGORYTMÓW SZTUCZNEJ INTELIGENCJI W UKŁADACH FPGA".

Z prośbą o przekazanie doktorantowi.

W niniejszym pliku przedstawiam zauważone drobne błędy edytorskie. Lokalizacja błędu podana jest w formacie: nr strony, nr wiersza od góry lub od dołu (np. s2g5 – strona 2 wiersz 5 od góry lub s9d8 – strona 9 wiersz 8 od dołu). Podana jest jedynie, część błędnego tekstu, bez propozycji korekty.

s17g6 inferencją

s19g11 rozwiązanie takio

s19d7 główne pod zadania

s21g15 Spratan

s21d13 model programowała

s23g3 Wiele zadań jest bardziej złożone

s23g16 doprowadził do powstania

s25d5 wzór matematyczny na obliczenia TF-IDF to..... (niedkończone)

s30g16 wycięty i n

s30g15 poniżej którego współczynników są wycinane

s31d14 słowinka

s31d4 czyto

s31d7 to model statystyczne

s33g4 śtatycznych"

s33g6 danych tekstowych czy szeregów (brak przecinka)

s33 wzór 2.13 – brak opisu zmiennej „y"

s33d10 wielu elementów sekwencji odległe (brak przecinka)

s33d6 . powoduje to (duża litera)

s34 brak opisów wzorów 2.14 – 2.19

s34d14 się gać (razem)

s35d19 wartości a na nich (brak przecinka)

s35d8 Jej zadanie (styl zdania)

s36d14 dopasowanie ale (brak przecinka)

s37d6 ... a konkretny (brak przecinka)

s37d5 .. a konkretną wartość błąd ...

s37 brak wyjaśnień wzorów 2.28 i 2.29.

s40d3 obsługuje inną częścią logiki

s46d1 formatowanie strony (tytuł podsekcji na następną stronę)

s47g2 serię eksperymentów sprawdzała schemat

s49g2 plus do hosta

s54g4 na znaczyć ilościach

s60d12 korelacja zachowana spadki

s74d2 (styl zdania/żargon)

s76g2 zamiast nr rysunku „wskoczyła" jego etykieta

s76g9 jak wyżej

s80g3 artykulików naukowych (może były jednak poważne naukowe?)

s80d7 Postawało

s80d8 ego problemu

s84d13 aby zaawansowaniu użytkownicy

s92g2 brakujący tekst (choć autor miał w planie)

s94 i 95 liczba Bernoullego -> Bernoulliego

s98g10 źa darmo

s102g6 pomiędzy feature mogą

s107d5 zapotrzebowanie za zasoby

s107d3 obliczenia które

s111d5 Operacje łączenia a następnie

s123g4 Oprócz day i godziny

s126g3 Wielkości

s126d3 a tyle niski żę nawet

s129d5 architektów sprzętowych

s134d1 Dzięki czystemu inteface'owi

s137g1 Przy zastosowani

Oczywiście, te drobne dostrzeżone literówki nie umniejszają wartości pracy. Mogą zostać wykorzystane przez Autora wg uznania.

Poznań, 30-11-2022