dr hab. inż. Dominik Belter, prof. PP
Politechnika Poznańska
Wydział Automatyki, Robotyki i Elektrotechniki

Poznań, 16 października 2024 r.

RECENZJA

rozprawy doktorskiej mgr inż. Dominiki Przewłockiej-Rus pt.: „Deep Neural Networks Quantization and Acceleration Methods for Real Time Energy Efficient Vision Systems"

1 Podstawa wykonania recenzji

Niniejsza recenzja dotyczy rozprawy doktorskiej mgr inż. Dominiki Przewłockiej-Rus zatytułowanej: „Deep Neural Networks Quantization and Acceleration Methods for Real Time Energy Efficient Vision Systems" w dyscyplinie naukowej automatyka, elektronika i elektrotechnika i technologie kosmiczne. Promotorem opiniowanej rozprawy doktorskiej jest prof. dr hab. inż. Marek Gorgoń, a promotorem pomocniczym dr inż. Tomasz Kryjak. Recenzja została przygotowana na podstawie uchwały Rady Naukowej Dyscypliny Automatyka, Elektronika, Elektrotechnika i Technologie Kosmiczne Akademii Górniczo-Hutniczej im. Stanisława Staszica w Krakowie nr RD AEEiTK/-510-6-4/24 z dnia 12.09.2024 r. Zgodnie z informacją dodatkową załączoną do powyższego pisma, postępowanie o nadanie stopnia naukowego doktora mgr inż. Dominice Przewłockiej-Rus procedowane na podstawie przepisów określonych przez Ustawę z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (Dz. U. 2023, poz. 742 z późniejszymi zmianami).

2 Rozprawa doktorska

2.1 Struktura rozprawy

Rozprawa doktorska zredagowana jest w formie cyklu 9 publikacji poprzedzonych wstępem i syntetycznym podsumowaniem osiągnięć i obejmuje 115 stron. Każdy z rozdziałów zawiera niezależną bibliografię. Dodatkowa bibliografia wykorzystana we wstępie i syntetycznym opisie pracy zawiera 26 pozycji.

Rozdział pierwszy wprowadza tematykę rozprawy i przedstawia motywację podjęcia tematu kwantyzacji i przyspieszania głębokich sieci neuronowych na urządzeniach brzegowych w celu uzyskania szybkich i wydajnych energetycznie systemów wizyjnych. Zauważony został wzrost wykorzystania zasobów związany z rozmiarami sieci neuronowych oraz zwiększone zużycie energii potrzebne do działania systemów opartych o te sieci. Poruszony został też problem działania nowych rozwiązań w czasie umożliwiającym szybkie przetwarzanie obrazu. Następnie przedstawiony jest problem badawczy, teza główna rozprawy, główny wkład Doktorantki w rozwój dyscypliny automatyka, elektronika i elektrotechnika i technologie kosmiczne oraz zawartość pracy.

W rozdziale drugim przedstawiono syntetyczny przegląd zawartości rozprawy będący podsumowaniem osiągnięć przedstawionych w cyklu 9 publikacji. W pierwszej kolejności przedstawione zostały metody zmniejszenia zajętości pamięci i przyspieszenia obliczeń, takie jak kwantyzacja parametrów czy logarytmiczna kwantyzacja do potęg dwójki przynoszące znaczne efekty na układach FPGAs, ASICs oraz powszechnie używanych procesorach CPU. Następnie pokazano główne wyniki będące efektem metody kwantyzacji do potęg dwójki oraz porównano je do wyników innych metod dostępnych w literaturze. Zaprezentowano również wyniki logarytmicznej i liniowej kwantyzacji, pokazując przewagę tej pierwszej. Podkreślono również znaczenie implementacji na dedykowanym sprzęcie, która pozwala na zastąpienie operacji mnożenia przesunięciem bitowym. Uwagę poświęcono również sprzętowej implementacji warstwy konwolucyjnej z odpowiednim kodowaniem wag i kwantyzacją PoT (ang. *Power-of-Two*). Zaproponowano metody przycinania i łączenia warstw, aby umożliwić dalsze zmniejszenie złożoności obliczeniowej i pamięci. Podsumowane zostały również eksperymenty otwartoźródłowych bibliotek Brevitas i FINN do optymalizacji czasu inferencji i zużycia energii na urządzeniach brzegowych. Wspomniane zostały również przykładowe zastosowania proponowanych rozwiązań, takie jak wykrywanie i śledzenie obiektów. W ostatniej części rozdziału podsumowane zostały główne osiągnięcia pracy.

Rozdział trzeci poświęcony został pracy pt. „XNOR CNNs in FPGA: real-time detection and classification of traffic signs in 4K - a demo" opublikowanej w materiałach konfrencji DASIP 2019: Conference on Design and Architectures for Signal and Image Processing. Praca prezentuje krótki opis implementacji na urządzeniu FPGA konwolucyjnej sieci neuronowej typu XNOR do rozpoznawania znaków na obrazach.

W rozdziale czwartym powiązanym z publikacją pod tytułem „Optimisation of a Siamese Neural Network for Real-Time Energy Efficient Object Tracking", przedstawiona jest autorska metoda śledzenia obiektów wykorzystująca efektywną implementację sieci bliźnia-

czych (ang. Siamese network) na urządzeniu FPGA. Przedstawiono efekty różnych metod kwantyzacji oraz wpływ przycinania nieistotnych połączeń w sieci. Jednocześnie przetestowano kwantyzację w trakcie uczenia i kwantyzację sieci po wyuczeniu. Poza zmniejszeniem rozmiaru sieci zaobserwowano również zwiększenie właściwości uogólniających (zmniejszenie ryzyka przeuczenia sieci), a także zauważono, że kwantyzacja w warstwach ukrytych zwiększa precyzję śledzenia.

W rozdziale piątym zademonstrowany został artykuł pt. „Quantised Siamese tracker for 4K/UltraHD video stream – a demo" dotyczący implementacji sieci bliźniaczych do śledzenia obiektów na układzie ZCU 104.

W rozdziale szóstym przedstawiono optymalizację działania sieci XNOR w zadaniu klasyfikacji znaków. Zaproponowany został akcelerator HDL dla sieci XNOR, który umożliwia wnioskowanie z prędkością prawie 450 klatek na sekundę. Jeszcze lepsze wyniki uzyskuje się dzięki drugiej metodzie - akceleratorowi Xilinx FINN - umożliwiającemu przetwarzanie obrazów wejściowych z szybkością około 550 klatek na sekundę. Oba podejścia zapewniają ponad 95% dokładność na zbiorze testowym.

Rozdział siódmy poświęcono publikacji powstałej we współpracy z Meta Reality Lab Research pt. „Power-of-two quantization for low bitwidth and hardware compliant neural networks", w której zaproponowano dwie metody kwantyzacji do potęg dwójki podczas trenowania sieci i zbadano ich wydajność. Przedstawiono również projekt sprzętowy, syntezę i analizę trzech cyfrowych mnożników dla kwantyzacji jednorodnej, wykazując najniższe zużycie zasobów i energii przy użyciu wag PoT oraz zademonstrowano oszczędności pamięci przy użyciu bardzo niskiej precyzji, co skutkuje kompresją rozmiaru sieci.

Rozdział ósmy dotyczy pracy pt. „Towards real-time and energy efficient Siamese tracking – a hardware-software approach" przedstawia implementację sprzętowo-programową sieci bliźniaczych do śledzenia obiektów na platformie Zynq UltraScale+ MPSoC ZCU104, ze szczegółową analizą czasową komponentów algorytmu i eksploracją przestrzeni projektowej pokazującą zależność między zużywaną mocą (wykorzystywanymi zasobami) a osiąganą prędkością (mierzoną w FPS). W pracy zaproponowany został algorytm-akcelerator, który pozwolił uzyskać dokładność śledzenia na równi z oryginalnym podejściem SiamFC, przy jednoczesnym znacznym zmniejszeniu liczby parametrów (a tym samym obliczeń).

Rozdział dziewiąty pokazuje, że sprzętowy akcelerator sieci neuronowej z wagami PoT zaimplementowany na układzie FPGA ZCU104 może być co najmniej 1,4 bardziej energooszczędny niż wersja z jednorodną kwantyzacją. W pracy pokazano projekt modułu Bitshift ACcumulate (BAC) implementującego filtrowanie konwolucyjne z wagami PoT, a więc z przesunięciem bitowym zamiast mnożenia, wraz z odpowiednim kodowaniem wag, implementację i porównanie sprzętowego akceleratora warstwy konwolucyjnej opartego na kwantyzacji jednorodnej i kwantyzacji logarytmicznej, w szczególności wykazującego znaczną redukcję zapotrzebowania na moc przy zastosowaniu wag PoT, również dla wysokich częstotliwości taktowania oraz algorytm przycinania dostosowany do kwantyzacji logarytmicznej, dzięki czemu można jeszcze bardziej zmniejszyć złożoność obliczeniową, a tym samym zapotrzebowanie na energię.

W rozdziale dziesiątym pt. „Power-of-Two Quantized YOLO Network for Pedestrian Detection with Dynamic Vision Sensor" zaproponowano architekturę, w której wagi warstw

splotu są kwantyzowane logarytmicznie do 4-bitowych wartości potęgi dwójki. Taka kompresja zmniejsza nie tylko ośmiokrotnie zużycie pamięci, ale także złożoność obliczeniową poprzez zastąpienie większości operacji mnożenia przesunięciem bitowym, ze względu na użycie wag będących potęgami dwójki. Jednocześnie proponowany system osiąga dokładność na równi ze zmiennoprzecinkową wartością bazową. Metoda została zaprezentowana w aplikacji detekcji obiektów, do których wykrywania została użyta architektura YOLO oraz dane z kamery zdarzeniowej.

W kolejnym rozdziale pt. „PowerYOLO: Mixed Precision Model for Hardware Efficient Automotive Detection with Event Data" zaroponowano wydajną sieć PowerYOLO o mieszanej precyzji INT8/LOG4 do wykrywania pieszych i pojazdów dla zbioru danych zdarzeń GEN1 cechującą się bardzo dobrym stosunkiem wydajności do złożoności obliczeniowej. Przedstawiono również metodę łączenia warstw splotu i normalizacji wsadowej w sposób uwzględniający specjalną postać wag sieci neuronowej, gdy są one sprowadzane do potęg dwójki.

2.2 Ogólna charakterystyka podjętej tematyki

Przedmiotem rozprawy jest optymalizacja szybkości działania, zużycia pamięci i energii przez sztuczne sieci neuronowe. Ze względu na potrzeby urządzeń działających w rzeczywistych scenariuszach inferencja z użyciem sieci neuronowych powinna być szybka. Jednocześnie z punktu widzenia urządzeń takich jak roboty mobilne czy pojazdy autonomiczne, ważne jest zużycie energii. Dlatego celowe jest opracowanie metod, które pozwolą na efektywne wykorzystanie sztucznych sieci neuronowych w urządzeniach mobilnych. W szczególności celem przeprowadzonych badań było zaproponowanie metod redukcji złożoności pamięciowej i obliczeniowej algorytmów, do rozmiaru odpowiedniego dla relatywnie niewielkich platform FPGA, przy jednoczesnym zachowaniu dokładności rozwiązania oraz spełnienia wymagań związanych z krótkim czasem obliczeń.

2.3 Problem naukowy i teza rozprawy

Problem naukowy i teza pracy zostały jednoznacznie zdefiniowane. Problemem naukowym rozpatrywanym w rozprawie jest efektywne wykorzystanie sztucznych sieci neuronowych na urządzeniach brzegowych. Celem przeprowadzonych badań było zaproponowanie metod redukcji złożoności pamięciowej i obliczeniowej algorytmów, do rozmiaru odpowiedniego dla relatywnie niewielkich zasobów urządzeń wbudowanych.

W pracy została zdefiniowana następująca teza:

Odpowiednie metody kwantyzacji parametrów sieci neuronowych pozwalają na znaczne zmniejszenie pamięci i złożoności obliczeniowej modeli, gwarantując jednocześnie zachowanie wysokiej dokładności i umożliwiając implementację systemów wizyjnych w czasie rzeczywistym na platformach sprzętowych o niskim opóźnieniu i niskim poborze mocy.

2.4 Rozwiązanie postawionego problemu

W celu udowodnienia tezy zaproponowano dwie metody uczenia sieci neuronowych Powers-of-Two (PoT) i uczenie poprzez wprowadzenie adaptacyjnego współczynnika uczenia w celu skompensowania nierównych odległości między poziomami kwantyzacji. Najlepsze wyniki osiągnięto dzięki metodzie opartej na podejściu STE (ang. *Straight Through Estimator*). Najpierw trenowany jest model w celu uzyskania precyzji sieci, a następnie szkolenie z uwzględnieniem kwantyzacji.

Zaproponowano również szereg eksperymentów wykorzystujących inne schematy kwantyzacji: tzw. radical binary quantisation (sieci neuronowe XNOR) oraz kwantyzację liniową. Szczególną uwagę poświęcono możliwościom akceleracji takich modeli w urządzeniach FPGA/SoC dla zaawansowanych systemów wizyjnych, pokazując jak takie aplikacje mogą być zaimplementowane i jakie zyski energetyczne wynikają z wyboru odpowiedniej platformy.

2.5 Oryginalny dorobek autora i jego znaczenie poznawcze i aplikacyjne

Oryginalny dorobek Doktorantki związany jest z analizą i rozwojem metod umożliwiających implementację funkcjonalnych systemów wizyjnych czasu rzeczywistego wykorzystujących algorytmy głębokich sieci neuronowych w urządzeniach o niskim poborze mocy i można go zdefiniować następująco:

- metody uczenia sieci neuronowych z wagami skwantowanymi do potęg dwójki (PoT), zapewniające dokładność porównywalną z modelami o pełnej precyzji, nawet w przypadku małej szerokości bitowej (4 bity), co jest trudne do osiągnięcia przy użyciu kwantyzacji liniowej.

- Zaprojektowanie architektury sprzętowej, która implementuje operację MAC dla sieci PoT przy użyciu przesunięcia bitowego zamiast mnożenia. Zaproponowany pojedynczy element obliczeniowy dla kwantyzacji 4×8 (szerokość wag × szerokość aktywacji) pozwala na dwukrotne zmniejszenie zapotrzebowania na energię w stosunku do modułu dla liniowo kwantyzowanego modelu o tej samej szerokości bitów.

- Zaprojektowanie architektury sprzętowej dla warstwy splotu PoT, która zużywa około 0,6 energii wymaganej dla warstwy standardowej i zwiększa zakres częstotliwości pracy.

- metoda fuzji warstw splotu i normalizacji wsadowej, uwzględniająca specjalną formę sieci neuronowej PoT.

- seria eksperymentów i analiza wpływu różnych schematów kwantyzacji (liniowych, logarytmicznych) z różnymi docelowymi szerokościami bitów, dla sieci neuronowych

stosowanych w zaawansowanych systemach wizyjnych, w celu oceny wpływu na dokładność, złożoność pamięciowo-obliczeniową i efektywność energetyczną takich systemów.

- projektowanie programowo-sprzętowych zaawansowanych i energooszczędnych systemów wizyjnych opartych na sieciach neuronowych do wykrywania znaków drogowych, śledzenia obiektów oraz wykrywanie pieszych i pojazdów.

  Zaproponowane metody kwantyzacji i ich odpowiednie wykorzystanie w celu zmniejszenia pamięci i złożoności obliczeniowej wybranych modeli sieci neuronowych, wraz z odpowiednią architekturą obliczeniową, umożliwiają implementację systemów wizyjnych czasu rzeczywistego w urządzeniach o niskim poborze mocy.

3 Uwagi merytoryczne

3.1 Silne i słabe strony pracy, uwagi dyskusyjne

- Co decyduje, że dana metoda nadaje się do implementacji na FPGA? Czy można wyciągnąć ogólne wnioski dotyczące cech sztucznych sieci neuronowych, które umożliwiają lub ułatwiają implementację i optymalizację na FPGA?

- Różnice pomiędzy różnymi metodami kwantyzacji przedstawionymi na rysunku 4 na stronie 39 nie są duże. Zaskakująco dobre wyniki osiąga sieć neuronowa dla kwantyzacji binarnej. Pojawia się pytanie, czy w sieciach neuronowych ważniejsze są wagi, czy liczba neuronów i połączeń między nimi. Jaki byłby wynik eksperymentu, który polegałby na zmniejszeniu liczby bitów reprezentujących wagi, ale jednoczesnym zwiększeniu liczby neuronów i połączeń między nimi, tak żeby zachować stały rozmiar sieci. Jaki to będzie miało wpływ na dokładność i szybkość działania?

- Dlaczego niektóre wyniki nie są dostępne w Tab. 2.1? Na przykład brakuje dokładności 4L/32F i 4L/8U dla DeepShift. Taka niepełna tabela utrudnia porównanie.

- Zdanie „Następnie czerwone i niebieskie piksele są segmentowane przy użyciu stałego progu" na stronie 28 jest niejasne. Czy oznacza to, że pierwsze dwa kanały (barwa i nasycenie) są używane do segmentacji?

3.2 Ważniejsze uwagi szczegółowe

- W wielu miejscach stosowany jest znak 'x' w znaczeniu wielokrotności. W latex jest do tego specjalny znak, który wstawiamy poleceniem „`\times`"

- Proponuję dodać oryginalne wyniki z sieci neuronowej bez kwantyzacji (wspomniane w tekście) do Tab. 3 na stronie 39, aby ułatwić porównanie wyników. Jaka jest różnica między wierszem "FP" w Tab. 3 a metodą z [2]?

- Jak na tle metod przedstawionych w tabeli 7 na stronie 54 wypada metoda z pracy [15]?
- Czym jest mAP50-90 gap w tabeli IV na stronie 111?
- W pracy można znaleźć kilka literówek, np. "computatnois",
- Pomiędzy liczbą a jednostką powinna być spacja. W wielu miejscach brakuje tej przerwy.

4 Wnioski końcowe

Powyższe uwagi nie umniejszają wartości merytorycznej pracy. Przyjęte w rozprawie założenia są prawidłowe, a Doktorantka rozwiązała postawione problemy naukowe oraz użyła właściwych metod. Rozprawa świadczy o wiedzy na zaawansowanym poziomie o charakterze podstawowym dla dziedziny nauk inżynieryjno-technicznych oraz o charakterze szczegółowym, dotyczącym metod redukcji złożoności pamięciowej i obliczeniowej algorytmów, do rozmiaru odpowiedniego dla relatywnie niewielkich platform FPGA. Rozprawa doktorska prezentuje ogólną wiedzę teoretyczną kandydatki w dyscyplinie oraz umiejętność samodzielnego prowadzenia pracy naukowej. W rozprawie przeprowadzono w sposób właściwy analizę źródeł, w tym literatury światowej dotyczącej metod kwantyzacji i optymalizacji sieci neuronowych. Mgr inż. Dominika Przewłocka-Rus wykazała się umiejętnością poprawnego i przekonującego przedstawiania uzyskanych przez siebie wyników. Wyniki przedstawione są w sposób zwięzły i przejrzysty. Na uwagę zasługuje również nawiązanie współpracy naukowej z Meta Reality Lab Research.

Przedstawiona do recenzji rozprawa doktorska stanowi oryginalne rozwiązanie aktualnego i ważnego problemu naukowego w dyscyplinie naukowej automatyka, elektronika, elektrotechnika i technologie kosmiczne oraz potwierdza wiedzę teoretyczną i umiejętność samodzielnego rozwiązywania problemów naukowych przez Doktorantkę. Rozprawa w pełni spełnia ustawowe wymagania dotyczące rozpraw doktorskich określone w artykule 187 ustawy z dnia 20 lipca 2018 r. - Prawo o szkolnictwie wyższym i nauce.

Wnioskuję do Rady Dyscypliny Automatyka, Elektronika i Elektrotechnika i Technologie Kosmiczne Akademii Górniczo-Hutniczej o dopuszczenie rozprawy mgr inż. Dominiki Przewłockiej-Rus do publicznej obrony.

Dominik Belter