Łódź, 04.02.2025

Dr hab. inż. Krzysztof Grudzień, prof. Politechniki Łódzkiej
Instytut Informatyki Stosowanej
Wydział Elektrotechniki, Elektroniki, Informatyki i Automatyki
Politechnika Łódzka
ul. Stefanowskiego 18
90-537 Łódź

RECENZJA

rozprawy doktorskiej mgr inż. Dominiki Przewłockiej-Rus

pt. *Metody kwantyzacji i akceleracji głębokich sieci neuronowych dla energooszczędnych systemów wizyjnych czasu rzeczywistego*

Recenzję przygotowałem w odpowiedzi na pismo Przewodniczącego Rady Dyscypliny Automatyki, Elektroniki, Elektrotechniki i Technologii Kosmicznych Akademii Górniczo-Hutniczej w Krakowie z dnia 12.09.2024 informującego mnie o powołaniu mojej osoby na recenzenta w przewodzie doktorskim mgr Dominiki Przewłockiej-Rus. Oceny rozprawy doktorskiej dokonano według kryteriów określonych w ustawie z 20 lipca 2018 roku Prawo o szkolnictwie wyższym i nauce. Promotorem rozprawy doktorskiej jest Profesor Marek Gorgoń, a promotorem pomocniczym Dr inż. Tomasz Kryjak.

1. Problematyka naukowa rozprawy

Rozprawa doktorska *„Metody kwantyzacji i akceleracji głębokich sieci neuronowych dla energooszczędnych systemów wizyjnych czasu rzeczywistego”* dotyczy istotnego zagadnienia związanego z badaniami rozwiązań energooszczędnych systemów wizyjnymi czasu rzeczywistego opartych na sieciach neuronowych oraz implementowanych w platformach sprzętowych. Przeprowadzone przez Doktorantkę prace badawcze wychodzą naprzeciw potrzebom dzisiejszych rozwiązań technologicznych. Rozwój systemów sztucznej inteligencji dedykowany analizie ogromnych ilości danych w czasie rzeczywistym jest jednym z kluczowych zagadnień świata nauki. W tym aspekcie należy również mieć na uwadze systemy wizyjne, które muszą spełniać ograniczenia energetyczne i sprzętowe. Obecnie w szerokorozumianych systemach wizyjnych stosuje się wiele typów sieci neuronowych, w tym konwolucyjne sieci neuronowe (CNNs, np. ResNet, EfficientNet), sieci syjamskie (Siamese Networks) (śledzenie obiektów), sieci neuronowe typu Transforme (ViTs) (wykrywanie obiektów), czy rekurencyjne sieci neuronowe RNN/LSTM (analiza sekwencji ruchu obiektów). Sieci te zazwyczaj działają operując na liczbach zmiennoprzecinkowych, przy wysokim zużyciu mocy obliczeniowej, skutkując znaczną energochłonnością. W tym aspekcie Doktorantka odnajduje zagadnienie opracowania systemów wizyjnych z wymogiem działania w czasie rzeczywistym przy założeniu

energoszędności i zachowaniu wystarczającego poziomu dokładności działania systemu. Efektywne wdrożenie rozwiązań AI w układach niskoenergetyczny systemów czasu rzeczywistego wymagało przeprowadzenie przez Doktorantkę prac badawczych zarówno na płaszczyźnie projektowania rozwiązań programistycznych, jak i docelowej implementacji sprzętowej. Stanowi to kolejne wyzwanie badawcze, które wymaga dogłębnego zrozumienia projektowanego rozwiązania. Opracowywane, w ramach rozprawy, metody kwantyzacji sieci neuronowych, mają na celu zmniejszenie wymagań sprzętowych poprzez redukcję rozmiaru modelu oraz minimalizację liczby operacji matematycznych przy jednoczesnym optymalnym wykorzystaniu zasobów sprzętowych. Dzięki zaproponowanym rozwiązaniom dane są przetwarzane z odpowiednią wydajnością, na urządzeniach o niskim zapotrzebowaniu energetycznym, gwarantując terminowe dostarczenie niezbędnych informacji do modułu kontroli lub nadzoru monitorowanego procesu. Przedstawiony w rozprawie problem badawczy został sformułowany prawidłowo, a poruszane przez Doktorantkę zagadnienie należy rozpatrywać jako bardzo istotne dla rozwoju nauki w dyscyplinie Automatyki, Elektroniki, Elektrotechniki i Technologii Kosmicznych.

2. Treść rozprawy – opinia, uwagi krytyczne i polemiczne

Recenzowana rozprawa doktorska, opierająca się na serii artykułów naukowych, przedstawia wyniki przeprowadzonych prac badawczych, w głównej mierze, dotyczących opracowania metod kwantyzacji wag sieci neuronowych oraz ich implementacji w układach o niskim poziomie poboru mocy. Mimo, że Doktorantka jasno podaje, że niniejsza rozprawa ma formę cyklu publikacji opisujących wyniki przeprowadzonych badań, nie jest to popularne podejście, aczkolwiek zgodne z wymogami, dla rozpraw doktorskich. Rozprawy składa się z 3 rozdziałów, w których Doktorantka wprowadza w tematykę zagadnienia, przedstawia cel rozprawy oraz hipotezy, opisuje uzyskane wyniki, dokonuje podsumowania prac badawczych oraz zamieszcza treści artykułów naukowych.

Celem badań było opracowywanie i przeanalizowanie wydajności metod kwantyzacji, umożliwiających implementację funkcjonalnych systemów wizyjnych czasu rzeczywistego z wykorzystaniem algorytmów głębokich sieci neuronowych w urządzeniach o niskim poborze mocy. W tym zakresie prace Doktorantki skupiły się nad redukcją złożoności pamięciowej i obliczeniowej algorytmów do rozmiarów odpowiednich dla stosunkowo kompaktowych platform sprzętowych przy jednoczesnym zachowaniu dokładności podstawowego rozwiązania (operacji na liczbach zmiennoprzecinkowych) oraz odpowiedniej organizacji obliczeń w celu spełnienia wymagań niskiego opóźnienia.

We wstępie rozprawy Doktorantka opisuje zagadnienie przetwarzania dużej ilości danych, sztucznej inteligencji oraz problemów spotykanych w systemach wizyjnych pojazdów autonomicznych. Doktorantka przedstawia charakterystykę układów wbudowanych, ze szczególnym uwzględnieniem układów FPGA (*Field-Programmable Gate Array*), pozwalających na programowalne przetwarzanie sygnałów i implementację układów cyfrowych, oferując elastyczność i wysoką wydajność, w tym akceleracją obliczeń dla sieci

neuronowych. Wartym podkreślenia jest fakt porównania głównych parametrów różnych platform sprzętowych. Choć są to istotne informacje, z punktu widzenia wyboru optymalnej platformy sprzętowej, to nie zostało wyjaśnione jak te parametry zostały wyznaczone. W przypadku dużych kart GPU wartości wydają się zaniżone. W tym aspekcie, bez wnikliwej analizy, przedstawiono charakterystykę rozważanych platform sprzętowych (FPGA, ASIC, SoC, eGPU, platformy neuromorficzne). Mimo, iż nie jest to główny cel rozprawy, z pewnością głębsze przedstawienie poszczególnych urządzeń jako potencjalnych platform sprzętowych do zastosowania przy prowadzonych pracach, wzbogaciłoby pracę o dodatkowe elementy.

W dalszej części wprowadzenia przedstawiony został rozwój głębokich sieci konwolucyjnych, jako podstawowych do analizy obrazów w systemach wizyjnych oraz coraz bardziej popularnych rozwiązań opartych na architekturze typu transformer (*ViT, Swin*). Podano również literaturę w zakresie prac badawczych nad algorytmami uczenia maszynowego, ze szczególnym uwzględnieniem i znaczeniem bazy *ImageNet* oraz modelu *AlexNet*. W rozprawie wspomniane zostały również rozwiązania oparte o sieci neuronowe dedykowane poszczególnym etapom i celom systemów wizyjnych; detekcji, śledzenia, segmentacji czy generowaniu obrazów (*YOLO, RCNNs, Siamese* networks, *GANs*, model *enkoder-dekoder*, model *dyfuzyjny*). Przedstawiony w rozprawie stan wiedzy, w tym zakresie, stanowi jedynie słowo wstępu do znaczącego obszaru badawczego jakim są systemy wizyjne, bez jego wnikliwej analizy. Mimo, że Doktorantka dalej podaje charakterystykę poszczególnych rozwiązań z uwzględnieniem liczby parametrów, dokładności, liczby operacji (GLOPS) oraz problemów jak nastręczały swoim autorom na drodze ich projektowania, w tym zakresie opis jest dość skromny. Mimo, że ta część tylko pokrótce oddaje znacznie i istotę stanu wiedzy w zakresie modeli AI dla systemów wizyjnych, wygląda znacznie lepiej niż ma to miejsce przy analizie dostępnych płaszczyzn sprzętowych. Rozdział *Intruduction* został skonstruowany przejrzyście i można stwierdzić, że zawiera wystarczającą ilość informacji, jednakże z punktu widzenia zawartości całej pracy należałoby się spodziewać więcej analizy i dyskusji nad opisywanymi kwestiami pod kątem stanu wiedzy. Zazwyczaj w artykułach naukowych, prezentujących oryginalne rozwiązania, brakuje miejsce na taką dyskusję.

W kolejnej sekcji rozprawy *Research problems, contributions and scientific novelty*, Doktorantka przedstawia hipotezę badawczą, która stwierdza, że (tłum. recenzenta) *odpowiednie metody kwantyzacji parametrów sieci neuronowych pozwalają na znaczną redukcję złożoności pamięciowej i obliczeniowej modeli, przy jednoczesnym zachowaniu wysokiej dokładności i umożliwieniu implementacji systemów wizyjnych czasu rzeczywistego na platformach sprzętowych o niskim opóźnieniu i niskim poborze mocy*. W dalszej części rozprawy, na bazie zaprezentowanych wyników w rozprawie doktorskiej i artkułach naukowych, należy potwierdzić jej słuszność. Doktorantka definiuje również wkład prac badawczych w rozwój dyscypliny Automatyki, Elektroniki, Elektrotechniki i Technologii Kosmicznych, zgodnie z dwoma następującymi zagadnieniami (tłum. recenzenta):

- *opracowanie metody trenowania sieci neuronowych kwantyzowanych do wag będących potęgami dwójki oraz projekt architektury sprzętowej uwzględniający*

specyficzną formę sieci, wraz z metodą fuzji warstw konwolucji i normalizacji wsadowej, uwzględniająca specyficzną formę sieci;

- przeprowadzenie serii eksperymentów i analizy wpływu różnych schematów kwantyzacji (liniowej, logarytmicznej) z różnymi docelowymi szerokościami bitowymi dla sieci neuronowych używanych w zaawansowanych systemach wizyjnych w urządzeniach wbudowanych, w celu oceny wpływu na dokładność, złożoność pamięciowo-obliczeniową i efektywność energetyczną takich systemów.

W przypadku drugiego zagadnienia należy zastanowić się nad jego sformułowaniem. Seria eksperymentów sama w sobie nie jest wkładem w rozwój dyscypliny. Stanowi zazwyczaj bazę do dalszych analiz i dyskusji, których to wynik może stanowić istotny wkład w rozwój dyscypliny. Brzmienie tego opisu wydaje się zaburzone, choć warto zauważyć, że w rozprawie znacząco podkreślono kwestię analizy wyników serii eksperymentów, co zdecydowanie podnosi poziom pracy.

Kluczowym rozdziałem rozprawy wydaje się rozdział *Synthetic overview of the dissertation*, który stanowi przegląd prac badawczych i uzyskanych wyników. Zgodnie z tą myślą, Doktorantka przedstawia zagadnienie redukcji złożoności pamięciowo-obliczeniowej modeli sieci neuronowych za pomocą kwantyzacji parametrów i aktywacji do wartości całkowitych. Szczególnie istotne dla prac Doktorantki są schematy kwantyzacji, do niskiej szerokości bitowej, które pozwalają na redukcję operacji MAC (operacja mnożenia i sumowania) do znacznie prostszych odpowiedników - w przypadku kwantyzacji logarytmicznej do potęg dwójki i przesunięć bitowych, a dla wag binarnych do operacji XNOR. Samo wprowadzenie do obszaru badawczego należy ocenić wystarczająco dobrze, jednak szersze spojrzenie z perspektywy stanu wiedzy, powołań na literaturę, mogłoby poprawić odbiór rozprawy. Zdecydowanie więcej informacji można odnaleźć w artykułach. Z pewnością dodanie do treści rozprawy analizy istniejących rozwiązań, w szerszym znaczeniu systemów wizyjnych np.: opartych o inne platforma obliczeniowa dla systemów AI, *edge computing* itp., pozwoliłoby uzyskać lepszy efekt znaczenia przeprowadzonych przez Doktorantkę prac naukowych.

Opracowane i zaprezentowane rozwiązania, trenowania sieci neuronowych uwzględniających efekt kwantyzacji *Quantization-Aware Training* (QAT), opierające się na metodach z kwantyzacją wag do potęg dwójki *Powers-of-Two* (PoT), zostały poddane analizie z zastosowaniem estymatora *Straight Through Estimator* ("STE") oraz adaptacyjnego współczynnika uczenia Adaptive Learning Rate (ALR) w celu kompensacji nierównych odległości między poziomami kwantyzacji. Opracowana przez Doktorantkę metoda (QAT) poddana została szerokim testom oraz analizie porównawczej z istniejącymi rozwiązaniami (m.in. *DeepShift*, *APoT*). W tej części rozprawy Doktorantka przedstawiła wyniki porównania działania procesu uczenia z uwzględnieniem kwantyzacji QAT, logarytmicznej i liniowej, dla odmiennych rozwiązań architektury głębokich sieci neuronowych z rodziny *Residual Networks* (ResNet). Zastosowanie dynamicznego zakresu pełnej skali (*Full Scale Range* - FSR) pozwoliło uzyskać poprawę wyników - lepsze dopasowanie do rozkładu wag. Istotą analizy porównawczej było przedstawienie efektywności działania opracowanego rozwiązania na tle

istniejących. W rozprawie można znaleźć szczątkowe informacje o procesie trenowania, jednakże opis tego zagadnienia jest odzwierciedlony w artykułach. W nich, przy znacznym udziale prac Doktorantki, opisano metodologię przeprowadzonych eksperymentów badawczych z zastosowaniem kwantyzacji logarytmicznej, pozwalającej zachować odpowiedni rozkład wag. Analiza uwzględniała również zastosowania tzw. współczynnika przycinania (*Pruning Factor*), a eksperymenty zostały przeprowadzone dla baz danych *CIFAR 10/CIFAR 100*, jak również na zbiorze *ImageNet*. Porównanie ze stanem wiedzy pokazało, że proponowane metody kwantyzacji osiągają dokładność zbliżoną do modeli zmiennoprzecinkowych, lub nawet je przewyższają, jak również okazały się skuteczniejsze od metod *DeepShift* oraz *APoT*. Ta część rozprawy doktorskiej warta jest podkreślenia pod kątem potwierdzenia skuteczności działania opracowanego rozwiązania. Przeprowadzone kompleksowe eksperymenty badawcze i wykonane analizy porównawcze stanowią dowód potwierdzający efektywność jego działania. Natomiast uzyskane wyniki wymagałyby szerszego komentarza ze strony Doktorantki; jakie argumenty przemawiają za tym, że zaproponowany model, mimo skwantyfikowanych wag, dostarcza wyniki zbliżone do tych uzyskanych przez modele oparte na wartościach zmiennoprzecinkowych?

Rozprawa Doktorska porusza również zagadnienie opracowania sprzętowej implementacji kwantyzacji wag warstw splotowych – moduł MAC dedykowany dla kwantyzacji PoT, nazywany BAC (Bitshift and Accumulate). Kolejny znaczny element, prowadzonych prac badawczych, to zaproponowany mechanizm przycinania wag (*death zone pruning*), który w odróżnieniu od klasycznego przycinania, przesuwa najmniejszy poziom kwantyzacji dalej od zera, nie zmniejszając liczby poziomów kwantyzacji. Wartym podkreślenia jest również fakt przeprowadzonej analizy zależności między częstotliwością zegara a konsumpcją mocy przez platformę sprzętową, co podnosi wiarygodność uzyskanych wyników w środowisku implementacyjnym.

Rozprawa przejawia również charakter aplikacyjny. Doktorantka przeprowadziła badania nad zastosowaniem opracowanych rozwiązań przy wykrywaniu pieszych oraz pojazdów. Przedstawiony model detekcji obiektów został zoptymalizowany pod kątem wydajności sprzętowej, wykorzystując mieszane precyzje obliczeń. Model ten jest przeznaczony do zastosowań w systemach wizyjnych o ograniczonym budżecie energetycznym, takich jak systemy wspomagania kierowcy, moduły pojazdów autonomicznych. Artykuł pokazuje, że kompleksowe podejście do projektowania systemów wizyjnych, uwzględniające zarówno algorytm, platformę sprzętową, jak i moduł sensorowy, może prowadzić do opracowania wydajnych i energooszczędnych rozwiązań. Mimo, że zaprezentowane rozwiązanie nie pokazuje istotnych postępów w samej jakości detekcji, np.: w porównaniu do sieci neuronowych typu Transformer, wskazuje jednak na znaczną oszczędność energii i sprzętową efektywności pracy systemów wbudowanych. Przedstawione rozwiązanie PowerYOLO, sieć o mieszanej precyzji, stanowi istotny wkład w rozwój energooszczędnej detekcji obiektów na urządzeniach wbudowanych. Przeprowadzona analiza porównawcza różnych konfiguracji rozwiązania pozwoliła na wyciągnięcie prawidłowych wniosków, choć w tym zakresie nie zweryfikowano, czy inne kombinacje precyzji mogłyby

zapewnić lepszy kompromis między dokładnością a oszczędnością zasobów. Wyniki pokazują, że różnice są niewielkie, co sugeruje, że warto byłoby zbadać jeszcze inne precyzje. Wartym podkreślenia jest fakt stosowania miary jakości detekcji obiektów z użyciem metryki mAP50-95, co ma kluczowe znaczenie przy dokładnej lokalizacji obiektów (np. w autonomicznych pojazdach, medycynie). Ta część rozprawy mogłaby zawierać dyskusję na temat wielkości różnic między wynikami testowanych rozwiązań, wraz z jej interpretacją, na ile te różnice mają znaczenie w praktycznym zastosowaniu opracowanych rozwiązań.

Rozprawa również odnosi się do początkowych prac badawczych Doktorantki i projektowania rozwiązań dla XNOR NN - binaryzacji jako jednej z najbardziej radykalnych form kwantyzacji, dedykowanych sprzętowej akceleracji obliczeń. Prace stanowiły bazę do dalszych badań nad przyspieszaniem binarnych sieci neuronowych w układach FPGA i ich integracji z systemami wizyjnymi czasu rzeczywistego. Element badawczy w tym zakresie stanowił zastosowanie sieci binarnych (XNOR) do klasyfikacji znaków drogowych oraz na przyspieszeniu ich działania za pomocą urządzeń FPGA. Prace Doktorantki uwzględniają również propozycję akcelerator *HDL* dla sieci XNOR. W tym zakresie przeprowadzono analizę porównawczą z istniejącym rozwiązaniem - akceleratorem *FINN* firmy *Xilinx*. Oba rozwiązania osiągają wysoką dokładność, a różnią się pod względem wydajności i zużycia energii. Akcelerator FINN wykazuje lepszą wydajność i niższe zużycie energii, jednak Doktorantka wskazuje, że opracowany akcelerator oferuje większą elastyczność w projektowaniu. Badania o charakterze aplikacyjnym wykazały możliwość implementacji syjamskich sieci neuronowej (*Siamese*) przeznaczonych do śledzenia obiektów w strumieniu wideo. Doktorantka skupia się na optymalizacji sieci neuronowej *Siamese* dla systemów wbudowanych (FPGA). Główne podejście polega na zmniejszeniu złożoności obliczeniowej poprzez kwantyzację i jej wpływ na efektywność energetyczną oraz precyzję operacji. Doktorantka wykonała analizę implementacji sprzętowo-programowej w pełni połączonej sieci *Siamese* (SiamFC), działającej w czasie rzeczywistym, na platformie sprzętowej (Zynq UltraScale+ MPSoC ZCU104), łączącej procesory ARM oraz programowalną logikę FPGA. Przeprowadziła optymalizację i eksplorację przestrzeni projektowej, aby znaleźć najlepszą równowagę między wydajnością a zużyciem energii na bazie metod kwantyzacji. Wyniki pokazały znaczną oszczędność energetyczną zaprojekowanego rozwiązania, przy bardzo dobrym wyniku wydajnościowym. To rozwiązanie zdecydowanie zmniejszyło konsumpcję energii w porównaniu do rozwiązań implementowanych na kartach graficznych GPU. Należy zauważyć tu bardzo interesującą analizę wpływu kwantyzacji na wydajność i jakość śledzenia oraz analizę efektywności energetycznej i szybkości obliczeń. Wartym podkreślenia jest tu fakt porównań z bezpośrednim zastosowaniem platform sprzętowych.

3. Pytania i uwagi do recenzowanej pracy

Recenzowana rozprawa doktorska podejmuje aktualny problem związany z opracowaniem rozwiązań opartych na sieciach neuronowych dedykowanych systemom wizyjnym z wymogiem działania w czasie rzeczywistym przy założeniu energoszędności

i zachowaniu wystarczającego poziomu dokładności. Praca została zaprezentowana w sposób dostępny i zrozumiały, a zbiór opublikowanych, przy znacznym udziale Doktorantki, i powiązanych tematycznie artykułów naukowych stanowił jej główną część. Sama treść rozprawy prowadzi czytelnika przez poszczególnego osiągnięcia Doktorantki zamieszczone w artykułach naukowych. Jednak podczas lektury rozprawy uwidacznia się brak głębszego i szerszego spojrzenia na zagadnienia podane przez Autorkę, szczególnie w kontekście stanu wiedzy. Pewne spostrzeżenia i pytania pojawiające się podczas lektury rozprawy doktorskiej nie powinny pozostać bez odpowiedzi ze strony Doktorantki, stąd poniższa lista pytań:

- jak należy rozumieć porównanie wyników zamieszczonych w tabeli 2.1, w kontekście zmiennej precyzji zapisu wartości liczbowych?
- jak należy interpretować wyniki o różnicy na poziomie dziesiątek czy setnych procenta? Na ile te różnice są znaczące?
- jakimi kryteriami kierowała się Doktorantka przy wyborze zbioru danych oraz typów sieci neuronowych przy testowaniu i weryfikacji opracowanych rozwiązań?
- czy w analizie rozwiązania były uwzględniane pełne koszty implementacji sprzętowej, dotyczące dodatkowej logiki przy implementacji kwantyzacji liniowej, logarytmicznej?
- czy zweryfikowano inne metody, niż podane w rozprawie, przycinania wag? Czy według Doktorantki może to mieć wpływ na wynik? Jeśli tak, to jaki?
- czy techniki regularyzacji/normalizacji (dropout/batch normalization) mogłyby pomóc w procesie kwantyzacji?
- czy w pracach badawczych uwzględniono analizę wpływu PoT na ramkę ograniczającą (*bounding boxy)*? Jaki może mieć to wpływ na wyniki?
- chociaż pseudo-obrazy są efektywną metodą przetwarzania danych z sensorów DVS, mogą one prowadzić do utraty pewnych informacji czasowych i polaryzacyjnych, które mogą być istotne dla dokładniejszej detekcji obiektów. Czy Doktorantka brała pod uwagę inny rodzaj danych? Czy wyniki mogą być bezpośrednio przenośne na inne zbiory danych lub inne typy zastosowanych sensorów? Czy użycie np.: histogramu zdarzeń (Hyper Histogram) czy transformacji czasowe (Temporal Active Focus, Stacking-Based on Time – SBT), spowodowałoby poprawę rezultatów?
- czy w pracach zidentyfikowano typ, rodzaj obrazów, dostarczających wyników o mniejszej czy większej dokładność? Czy istnieją reguły określające, w jakich warunkach, przy jakiej zawartości i jakości obrazu opracowane rozwiązanie sprawdza się lepiej lub gorzej?
- czy testowane były inne wartości szerokości bitowej niż te wspomniane w pracy?
- czy i jak zoptymalizowano transfer danych między FPGA i CPU?
- czy jest możliwości efektywnej implementacji architektury sieci neuronowych Transformer w układach FPGA?
- jak wypadłoby porównanie opracowanego rozwiązania do takich rozwiązań jak ZeroQuant czy LSSQ (Learned Step Size Quantization)?

- czy i jak opracowane rozwiązanie, po zaproponowanych modyfikacjach (?), mogłoby skutecznie realizować zadania sekwencyjne, np. przetwarzanie mowy czy przetwarzanie języka naturalnego (NLP)?
- czy opracowany schemat fuzji, warstwy konwolucji i normalizacji wsadowej, może skomplikować implementację modelu na różnych platformach sprzętowych, wymagając dodatkowych modyfikacji i optymalizacji?
- czy porównanie do innych platform, np.: NVIDIA Jetson, pozwoliłoby również uzyskać przewagę opracowanych w ramach rozprawy doktorskiej rozwiązań, jak przedstawiono w artykułach?

4. Ocena redakcji i przygotowania rozprawy

Praca została przedstawiona w sposób przejrzysty i nie budzący krytycznych zasadniczych uwag. Można jednak wskazać kilka kwestii, które od strony redakcji i przygotowania rozprawy mogą być istotne. Zastanawiające jest dobranie proporcji między wstępem do pracy, a syntetycznym przeglądem rozprawy. Oczywistym jest fakt, że rozprawa doktorska opiera się na cyklu publikacji naukowych. Jednakże w ocenie recenzenta syntetyczny przegląd powinien obejmować znaczną część treści. Z drugiej strony, jak wspomniano w rozdziale drugim recenzji, niewystraczająco rozwinięty został stan wiedzy - charakterystyka kontekstu prowadzonych badań. Dodatkowo, umieszczenie w rozprawie schematu obrazującego wzajemne relacje pomiędzy poszczególnymi opracowanymi rozwiązaniami oraz ich powiązania ze stanem wiedzy, w tym także ich zastosowania w praktycznych aspektach, ułatwiłoby ocenę rozprawy oraz zrozumienie procesu badawczego.

Jako uwagę redakcyjną należy rozpatrzyć pozycje zamieszczonych ilustracji/rysunków/tabel w relacji do odwołania się do nich w treści rozprawy, np.: w przypadku *Figure 1.1*, odwołanie do wykresu (zależności liczby parametrów modelu od dokładność) pojawia się w treści rozprawy po wykresie. Podobnie ma się rzecz z odwołaniem do tabeli 2.1, 2.3. Pojawiają się również braki powołań na rys. 2.2, 2.3 w treści rozprawy. W przypadku ilustracji 1.1, czytelnik spodziewa się odnośnika do źródła wykresu w podpisie pod rysunkiem.

Dane, wartości liczbowe w tabeli 2.1 są prezentowane z różną precyzją zapisu, a porównanie do modelu floating-point z użyciem trzech miejsc po przecinku powoduje niezrozumienie tego zapisu (tu DeepShift).

Dodatkowo, wydaje się, że zaprezentowane dane w Tabeli 2.3 są w pewnym stopniu nadmiarowe. Wyróżnienie zapisu *„**Overview of the scope of the research with key results** is presented below."* – zapis pogrubioną czcionką, można uznać za niepotrzebny.

Należałoby również się zastanowić, czy zamieszczenie artykułów powinno być uwzględnione w rozprawie jako kolejny rozdział nr 3. Czy w takim wypadku, wnioski nie powinny się znaleźć na końcu rozprawy? W spisie treści powinien się pojawić się również punkt dotyczący spisu literatury. Jak zostało wspomniane wcześniej, literatura zamieszczona w rozprawie jest dość skromna.

5. Wnioski końcowe

Recenzowana rozprawa, a w szczególności wyniki prac Doktorantki przedstawione w artykułach naukowych, prezentują istotne wyniki dla rozwoju dyscypliny Automatyki, Elektroniki, Elektrotechniki i Technologii Kosmicznych. Dotyczą one prac badawczych nad schematami kwantyzacji dostosowanymi do rozkładu wag w warstwach sieci neuronowych, przeznaczonych do systemów wizyjnych czasu rzeczywistego, implementowanych w urządzeniach wbudowanych o niskim poborze mocy. W pracach Doktorantki wykazano, że opracowana metoda kwantyzacji dla małych szerokości bitowych utrzymuje dokładność, dla zastosowanego zbioru danych obrazowych, na równi z modelami pełnej precyzji i przewyższa wyniki osiągane z zastosowaniem istniejących metod kwantyzacji. Wartym wskazania jest również fakt przeprowadzenia licznych eksperymentów badawczych, pozwalających na bogatą ewaluację doświadczalną opracowanych rozwiązań. Należy również podkreślić aplikacyjny charakter pracy w zakresie projektowania systemów wizyjnych, łączących płaszczyznę oprogramowania i sprzętu, z zastosowaniem sieci neuronowych do wykrywania znaków drogowych, pieszych i pojazdów oaz śledzenia obiektów. Zaproponowane metody kwantyzacji i ich odpowiednie zastosowanie, w celu zmniejszenia złożoności obliczeniowej i wymaganej pamięci w wybranych modelach sieci neuronowych, wraz z odpowiednią architekturą obliczeniową, umożliwiają implementację systemów wizyjnych czasu rzeczywistego w urządzeniach o niskim poborze mocy, co dowodzi słuszności tezy.

Podsumowując, rozprawa doktorska mgr inż. Dominiki Przewłockiej-Rus stanowi oryginalne rozwiązanie problemu naukowego i wskazuje na wysoki poziom ogólnej wiedzy teoretycznej Autorki z zakresu dyscypliny Automatyki, Elektroniki, Elektrotechniki i Technologii Kosmicznych. Doktorantka w swojej rozprawie doktorskiej podejmuje aktualny problem badawczy. Sformułowane przez Doktorantkę wnioski, należy uznać za trafne i ważne z punktu widzenia rozwoju metod kwantyzacji i ich odpowiedniego zastosowania. Doktorantka, na podstawie przedstawionego w pracy wkładu w poszczególne artykuły naukowe, wykazała się umiejętnością samodzielnego rozwiązania problemu badawczego. Mimo uwag i komentarzy zawartych w recenzji, rozprawa doktorska pozostaje na bardzo wysokim poziomie i nie ma to wpływu na jej ogólną ocenę. W razie spełnienia kryteriów ustalonych przez Radę Dyscypliny Automatyki, Elektroniki, Elektrotechniki i Technologii Kosmicznych Akademii Górniczo-Hutniczej w Krakowie, rekomenduję ją do rozważenia jako wyróżniającą się rozprawę doktorską.

Stwierdzam, że opiniowana rozprawa doktorska spełnia wymogi ustawy z dnia 20 lipca 2018 Prawo o szkolnictwie wyższym i nauce. Wnoszę o dopuszczenie mgr inż. Dominiki Przewłockiej-Rus do publicznej obrony pracy doktorskiej.

Signed by / Podpisano przez
Krzysztof Grudzi
Politechnika Łódzka
Date / Data:
2025-02-06 09:0