AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

DZIEDZINA NAUK INŻYNIERYJNO-TECHNICZNYCH

DYSCYPLINA AUTOMATYKA, ELEKTRONIKA, ELEKTROTECHNIKA I TECHNOLOGIE KOSMICZNE

AUTOREFERAT ROZPRAWY DOKTORSKIEJ

METODY KWANTYZACJI I AKCELERACJI GŁĘBOKICH SIECI NEURONOWYCH DLA ENERGOOSZCZĘDNYCH SYSTEMÓW WIZYJNYCH CZASU RZECZYWISTEGO

AUTORKA: DOMINIKA PRZEWŁOCKA-RUS

PROMOTOR ROZPRAWY: PROF. DR HAB. INŻ. MAREK GORGOŃ

PROMOTOR POMOCNICZY: DR INŻ. TOMASZ KRYJAK

PRACA WYKONANA: AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE, WYDZIAŁ ELEKTROTECHNIKI, AUTOMATYKI, INFORMATYKI I INŻYNIERII BIOMEDYCZNEJ, KATEDRA AUTOMATYKI I ROBOTYKI

KRAKÓW 2025

Streszczenie

Projektowanie energooszczędnych systemów wizyjnych o małej latencji i wysokiej skuteczności działania wymaga podejścia charakteryzującego się spójnością między rozwiązaniem algorytmicznym i docelową platformą, na której zostanie ono uruchomione. Głównym wyzwaniem jest implementacja złożonych pamięciowo-obliczeniowo sieci neuronowych w niewielkich urządzeniach małej mocy, jak SoC (System on Chip), FPGA (Field-Programmable Gate Array) czy docelowo ASIC (Application-Specific Integrated Circuit). Stosuje się zatem szereg metod pozwalających na redukcję tej złożoności, poprzez zmniejszenie rozmiarów modelu lub uproszczenie obliczeń, w szczególności operacji mnożąco-akumulujących: jedną z nich jest kwantyzacja parametrów sieci do liczb całkowitych. Pewnym standardem stała się kwantyzacja liniowa 8-bitowa, natomiast to inne, specjalne schematy pozwolą na realizację zaawansowanych systemów o znacznie wyższej wydajności. W ramach przeprowadzonych badań zaproponowano metody uczenia i wydajnej implementacji sprzętowej sieci kwantyzowanych do wag o wartościach potęg dwójki, pozwalając na realizację modeli 4-bitowych o skuteczności porównywalnej do sieci pełnej precyzji i jednocześnie umożliwiając znaczną redukcję złożoności obliczeniowej poprzez zmianę mnożenia na operację przesunięcia bitowego. Ponadto przeanalizowano też możliwości użycia różnych schematów kwantyzacji (liniowej, logarytmicznej, binarnej i mieszanej) dla sieci neuronowych używanych w zaawansowanych systemach wizyjnych, proponując modele dedykowane urządzeniom niewielkiej mocy, dla zadań klasyfikacji, śledzenia oraz detekcji obiektów. W wyniku odpowiedniego projektowania algorytmów i ich implementacji w platformach wbudowanych pokazano, że odpowiednie metody kwantyzacji modeli umożliwiają realizację systemów o wysokiej skuteczności działania, małej latencji i niskim poborze energii.

Rozprawa doktorska stanowi cykl dziewięciu publikacji dotyczących metod kwantyzacji i akceleracji sieci neuronowych dla szczególnego problemu energooszczędnych systemów wizyjnych działających w czasie rzeczywistym. Rozprawa zawiera także dwa dodatkowe rozdziały: wstęp wprowadzający do problematyki badawczej wraz z opisem wkładu autorki w dyscyplinę, oraz syntetyczne omówienie rozprawy przedstawiające najważniejsze wyniki.

Motywacja

Zastosowanie systemów sztucznej inteligencji w systemach autonomicznych mających przejąć pewną odpowiedzialność człowieka wiąże się z szeregiem wyzwań. Oczywistym przykładem takich systemów są autonomiczne pojazdy (wojskowe drony czy samochody), jak również zaawansowane systemy wspomagania kierowcy. Autonomiczne pojazdy często wyposażane są w kamery i inne czujniki - różnego typu radary czy sensory termowizyjne i ultradźwiękowe, stanowiące system percepcji pojazdu dostarczający niezbędnych informacji o otoczeniu. Co szczególnie istotne, dane zbierane z niektórych innych niż kamera czujników również mogą być analizowane z użyciem algorytmów przetwarzania obrazów cyfrowych (lub podobnych): na przykład z czujników LiDAR generujących chmurę punktów 3D czy kamer zdarzeniowych rejestrujących chmurę zdarzeń. Odpowiednia analiza zbieranych danych pozwala na podjęcie właściwych decyzji sterujących pojazdem. W szczególności znajdą tu zastosowanie wszelkie algorytmy wykorzystujące sieci neuronowe (głównie konwolucyjne, ale również grafowe czy tzw. transformersy), wyspecjalizowane do szukania i odtwarzania wzorców w zadaniach detekcji, śledzenia czy segmentacji. Sam system sterowania może być zaprojektowany *standardowo*, tzn. algorytmy będą uczone *offline* na ogromnych ilościach danych reprezentujących rzeczywiste warunki, lub korzystając z algorytmów uczenia ze wzmocnieniem, gdzie agent (autonomiczny pojazd) w pewien pseudo-losowy sposób eksploruje środowisko i z pomocą sygnału nagrody uczy się, które akcje są opłacalne. Jednocześnie, w rozważanych aplikacjach konieczne jest działanie z niską latencją, tzn. odpowiedź systemu musi być otrzymana z minimalnym opóźnieniem, najlepiej mniejszym niż czas potrzebny na reakcję człowieka w podobnej sytuacji. Zwyczajowo taki system określać bę-

dzie się jako działający w czasie rzeczywistym. Dodatkowo konieczna jest gwarancja wysokiej precyzji działania - system musi podejmować decyzje poprawne i logiczne z punktu widzenia człowieka, i musi to robić co najmniej tak często jak człowiek. W końcu, całość powinna działać w urządzeniach o ograniczonym budżecie energetycznym, często zasilanych akumulatorowo. Równoczesne spełnienie warunków działania w czasie rzeczywistym oraz wysokiej skuteczności wymaga odpowiedniego podejścia do przeprowadzenia obliczeń – w szczególności wykorzystania wielowątkowości (np. wielu CPU lub GPU). Dodając trzeci wymóg – energooszczędności – zdefiniuje się problem, który wymusza zmianę podejścia do tworzenia algorytmów z paradygmatu rozwiązań o rozmiarach liczonych w gigabajtach (*duże* sieci neuronowe, które stanowią podstawę najbardziej ekscytujących społeczeństwo aplikacji) do algorytmów kompaktowych, uruchamianych w urządzeniach wbudowanych o niskiej mocy.

Cele i problemy badawcze

Celem prowadzonych badań była analiza i opracowanie metod umożliwiających implementację funkcjonalnych systemów wizyjnych czasu rzeczywistego używających algorytmów głębokich sieci neuronowych w urządzeniach niskiej mocy. W szczególności oznaczało to zaproponowanie metod redukujących złożoność pamięciowo-obliczeniową użytych algorytmów do rozmiarów odpowiednich dla relatywnie niewielkich platform FPGA (lub docelowo dedykowanych układów), przy jednoczesnym założeniu utrzymania skuteczności działania rozwiązań bazowych (bez ograniczeń precyzji obliczeń), oraz użycie odpowiedniej orkiestracji obliczeń (wykonywanych równolegle) w celu spełnienia warunku niskiej latencji systemu.

Główny wkład autorki w dyscyplinę *automatyki, elektroniki, elektrotechniki i technologii kosmicznych* w ramach prowadzonych badań można podsumować w następujących punktach:

1. Zaproponowanie metod uczenia sieci kwantyzowanych do wag o wartościach potęg dwójki, zaprojektowanie architektury sprzętowej realizującej operację mnożenia i akumulacji (oraz warstwę konwolucyjną), z uwzględnieniem szczególnej postaci takiej sieci, oraz metody fuzji warstw konwolucyjnej i normalizacji partiami, uwzględniającą szczególną postać takiej sieci.

2. Przeprowadzenie szeregu eksperymentów i analizy wpływu różnych schematów kwantyzacji (liniowej, logarytmicznej), z różną docelową szerokością bitową, dla sieci neuronowych używanych w zaawansowanych systemach wizyjnych, celem oceny wpływu na skuteczność, złożoność pamięciowo-obliczeniową oraz energooszczędność takich systemów.

Zaproponowana seria publikacji dotyczy badań prowadzonych w odniesieniu do wiodącej hipotezy badawczej: ***odpowiednie metody kwantyzacji parametrów sieci neuronowych pozwalają na znaczną redukcję złożoności pamięciowo-obliczeniowej modeli, jednocześnie gwarantując zachowanie wysokiej skuteczności działania i umożliwiając realizację w platformach sprzętowych z niską latencją i niskim poborem energii.***

Najważniejsze wyniki

W ramach prowadzonych badań zaproponowano dwie metody uczenia sieci dwójkowych PoT (*Power-of-Two*), przy czym najlepsze wyniki osiągnięto z użyciem metody czerpiącej z podejścia *Straight Through Estimator*. W pierwszej kolejności uczy się model pełnej precyzji, a następnie przeprowadza uczenie kwantyzowane realizując przejście w przód z użyciem wag kwantyzowanych do wartości potęg dwójki (np. 2^0, 2^{-1}, itp.) z zadaną szerokością bitową. Następnie propagacja wsteczna realizowana jest na liczbach zmiennoprzecinkowych i po aktualizacji wag, model jest na nowo kwantyzowany (a zatem kwantyzacja przeprowadzana jest między kolejnymi iteracjami).

Tabela 1: Wyniki dla modelu ResNet18 uczonego na zbiorze ImageNet, dla różnych metod kwantyzacji (np. 4L/8U oznacza kwantyzację warstw konwolucyjnych (C) logarytmicznie (L) do szerokości 4-bitowej i kwantyzację warstw w pełni połączonych (FC) z użyciem kwantyzacji liniowej (U) 8-bitowej), w porównaniu do innych metod PoT SOTA (*State-Of-The-Art*). W nawiasach podano różnicę w skuteczności klasyfikacji względem modelu zmiennoprzecinkowego

Model	Precyzja (C/FC)	Zaproponowana	DeepShift	APoT
ResNet18	4L/32F	70.0% (+0.2)	-	70.7% (+0.5)
ResNet18	4L/8U	69.9% (+0.1)	-	-
ResNet18	4L/4L	69.5% (-0.2)	69.6% (-0.2)	-

Tabela 2: Porównanie modeli kwantyzowanych logarytmicznie (PoT) oraz liniowo do szerokości 4-bitowej

Model	FP32	PoT	Liniowa
ResNet20 CIFAR 10	91.8%	91.6% (-0.17)	91.22% (-0.55)
ResNet20 CIFAR 100	68.7%	68.5% (-0.2)	65.5% (-3.2)
ResNet18 ImageNet	69.8%	69.9% (+0.1)	57.8% (-10.9)

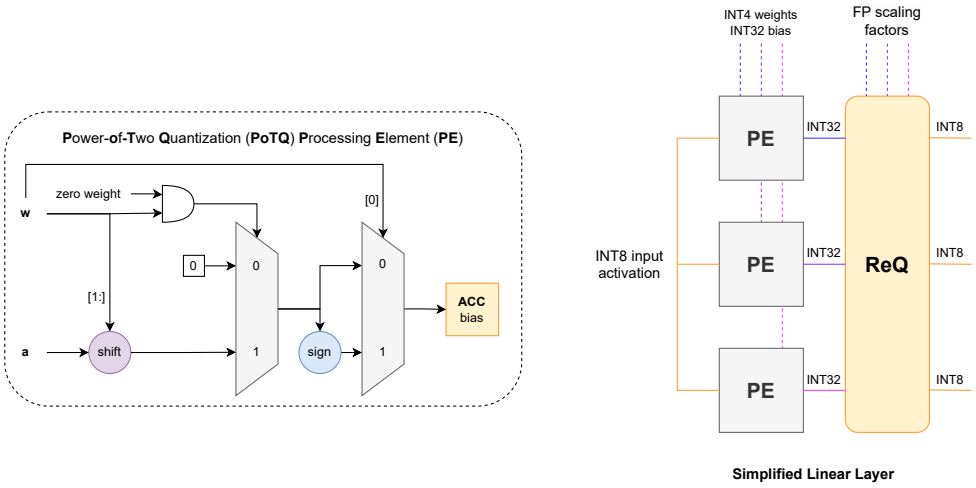
Metodę przetestowano dla wielu architektur i porównano z rozwiązaniami SOTA wykazując, że zaproponowana metoda pozwala osiągnąć wyniki porównywalne z innymi metodami korzystającymi z wag PoT i wariacji wag PoT - APoT (wagi są sumą wartości PoT). Reprezentatywne wyniki przedstawiono w tabeli 1.

Ponadto, przeprowadzono też szereg eksperymentów porównujących modele 4-bitowe kwantyzowane logarytmicznie i liniowo, pokazując przewagę kwantyzacji PoT - skrót wyników przedstawiono w tabeli 2 -, oraz dowodzących możliwości wprowadzenia znacznych uproszczeń w kwestii złożoności pamięciowo-obliczeniowej, w tym w kontekście urządzeń wbudowanych. Pokazano, że dla zadania klasyfikacji wprowadzenie wag PoT 4-bitowych pozwala na kompresję wag niemalże dwukrotnie większą niż w przypadku klasycznej kwantyzacji 8-bitowej, przy jednoczesnym utrzymaniu skuteczności sieci 8-bitowej (czy nawet zmiennoprzecinkowej 32-bitowej).

Zaprojektowano również dedykowany moduł sprzętowy MAC, w którym operacja mnożenia zastąpiona jest wydajną operacją przesunięcia bitowego, co jest możliwe dzięki użyciu specyficznego schematu kwantyzacji PoT (architektura elementu obliczeniowego pokazana jest w lewej części rysunku 1). Porównano taki element obliczeniowy dla wag 4-bitowych i aktywacji 8-bitowych dla kwantyzacji liniowej, PoT, APoT, oraz ze standardowym elementem obliczeniowym dla modelu z wagami i aktywacjami kwantyzowanymi liniowo do szerokości 8-bitowej. Tym samym pokazano, że moduł dla 4-bitowej kwantyzacji PoT używa najmniejszą liczbę elementów elektronicznych, co przekłada się również na kilkukrotnie niższe zapotrzebowanie energetyczne: x6 dla PoT względem modułu MAC 8x8 dla kwantyzacji liniowej oraz x2 względem modułu MAC 4x8 dla kwantyzacji liniowej. Wy-

czerpujący opis zaproponowanych metod uczenia, implementacji sprzętowych i więcej wyników przedstawiono w [Przewłocka-Rus et al., 2022].

Następnie zaproponowano sprzętową realizację warstwy konwolucyjnej PoT (schemat w prawej części rysunku 1), z odpowiednim kodowaniem wag, na platformie Zynq UltraScale+ MPSoC ZCU104. Wykazano, że



Rysunek 1: Schemat uproszczonej kwantyzowanej warstwy sieci neuronowej z elementem BAC dla kwantyzacji logarytmicznej PoT. Po odczytaniu z pamięci odpowiednich wag i biasów, elementy przetwarzają aktywacje z poprzedniej warstwy korzystając z wydajnej operacji przesunięcia bitowego. Aktywacje wyjściowe są następnie rekwantyzowane z użyciem odpowiednich współczynników skalujących.

warstwa używająca dedykowanego kwantyzacji PoT modułu MAC - zwanego BAC (ang. *Bitshift and Accumulate*) - używa ok. 0.6 energii potrzebnej dla warstwy ze standardowymi modułami MAC, dla aktywacji 8-bitowych i wag 4-bitowych, jednocześnie zwiększając zakres możliwej wysokości częstotliwości pracy układu.

Aby rozszerzyć uniwersalność kwantyzacji PoT względem możliwości gwarantowanych przy użyciu kwantyzacji liniowej, zaproponowano metody wygaszania połączeń oraz fuzji warstw, umożliwiające dalsze redukcje złożoności pamięciowo-obliczeniowej.

Kwantyzacja liniowa dla niskich szerokości bitowych wprowadza automatyczne wygaszenie połączeń o wartościach mniejszych niż najniższy poziom kwantyzacji, bez wpływu na ostateczną liczbę poziomów kwantyzacji. Kwantyzacja logarytmiczna nie posiada takiej właściwości, a bezpośrednie zerowanie wag o najniższych wartościach prowadzi do zmniejszenia liczby przedziałów kwantyzacji (co może mieć istotny wpływ na skuteczność rozwiązania). Aby umożliwić poprawne wygaszanie wag dla kwantyzacji logarytmicznej, zaproponowano więc metodę podwójnej normalizacji, wprowadzającą *martwą strefę* o zadanej szerokości, odsuwając najmniejszy przedział kwantyzacji dalej od zera. Dzięki temu zabiegowi możliwe jest wygaszanie połączeń analogicznie jak dla kwantyzacji liniowej: bez straty liczby poziomów kwantyzacji, i, jak pokazano w eksperymentach, dla sieci redundantnych, jak ResNet20 dla zbioru CIFAR10, możliwe jest wygaszenie ponad 40% połączeń bez straty na skuteczności działania sieci (przy wygaszeniu 70% połączeń strata wyniosła niecały punkt procentowy). Szczegółowy opis metody wygaszania połączeń, jak również akceleracji warstw konwolucyjnych z użyciem modułu BAC przedstawiono w [Przewłocka-Rus and Kryjak, 2022a].

W celu redukcji liczby obliczeń w trakcie inferencji standardowo przeprowadza się również fuzję warstw konwolucyjnych z warstwami normalizacji partiami, odpowiednio modyfikując wagi i *bias* warstwy konwolucyjnej o współczynniki warstwy normalizacji partiami. Naturalnie taka modyfikacja nie wpływa na skuteczność sieci neuronowej, ale istotnie redukuje liczbę koniecznych operacji mnożenia i dodawania, oraz liczbę parametrów.

Wprowadzenie dokładnie takiej fuzji do sieci dwójkowej skutkowałoby przekształceniem wag o wartościach potęg dwójki do wartości *dowolnych* rzeczywistych, tracąc tym samym istotną właściwość umożliwiającą użycie elementów obliczeniowych BAC. W celu redukcji liczby operacji proponuje się zatem wprowadzenie dwóch różnych zabiegów prowadzących do tożsamych uproszczeń. Ze względu na to, że bias nie jest poddawany kwantyzacji logarytmicznej, z powodzeniem można dokonać jego modyfikacji zgodnie ze schematem znanym ze standardowej fuzji warstw konwolucyjnej i normalizacji partiami. Mnożnik związany z wagą jest za to łączony ze współczynnikiem skalującym operacji kwantyzacji. W ten sposób redukuje się maksymalnie wszystkie dodatkowe obliczenia związane z warstwą normalizacji partiami, przy jednoczesnym jedynie niewielkim wzroście złożoności pamięciowej względem modelu po kompletnej fuzji - zamiast jednego współczynnika skalującego dla całej warstwy otrzymuje się współczynniki skalujące dla każdej mapy wyjściowej osobno.

Całość zaproponowanych metod pozwala na projektowanie wydajnych systemów wizyjnych o SOTA stosunku złożoności modelu i architektury obliczeniowej do skuteczności rozwiązania, jak pokazano na przykładzie modelu mieszanej precyzji (wagi 4-bitowe PoT, aktywacje 8-bitowe kwantyzowane liniowo) PowerYOLO dla detekcji pieszych i pojazdów osiągając skuteczność mAP50-95 0.301 (8.3% spadku względem modelu bazowego, bez redukcji precyzji obliczeń), przy jednoczesnej redukcji rozmiaru 8x i wprowadzając znaczne uproszczenia w architekturze obliczeniowej, poprzez zmianę operacji mnożenia na operację przesunięcia bitowego. Szczegółowy opis metod i wyniki porównawcze zaprezentowano w [Przewłocka-Rus and Kryjak, 2023] oraz [Przewłocka-Rus et al., 2024].

W ramach prowadzonych badań zaproponowano też szereg eksperymentów z użyciem innych schematów kwantyzacji: radykalnej kwantyzacji binarnej (do postaci sieci XNOR) oraz kwantyzacji liniowej. W tej części szczególną uwagę zwrócono na możliwości akceleracji takich modeli w urządzeniach FPGA/SoC dla zaawansowanych systemów wizyjnych, pokazując sposoby implementacji takich aplikacji oraz zyski energetyczne będące konsekwencją wyboru odpowiedniej platformy.

Na potrzeby systemu detekcji znaków drogowych zaprojektowano sprzętowy akcelerator sieci neuronowych XNOR, z aktywacjami i wagami w binarnej postaci, i tym samym umożliwiający zredukowanie operacji konwolucji do kilku prostych operacji bitowych. Akcelerator został zaimplementowany semi-potokowo: obliczenia w każdej warstwie konwolucyjnej zrównoleglone zostały względem filtrów, a mapy wyjściowe zapisywane są do dedykowanych pamięci BRAM. Dla zaproponowanej architektury sieci binarnej (zbliżonej do sieci LeNet5 i osiągającej skuteczność 96.28% na zbiorze GTSRB), dla wejścia 32x32, uruchomiony na pltaformie Zynq UltraScale+ MPSoC ZCU104 akcelerator pozwolił na przetwarzanie prawie 450 FPS przy poborze energii 4.396W.

Przeprowadzono również eksperymenty porównawcze dla systemu detekcji znaków drogowych z użyciem nowych narzędzi Brevitas i FINN, pozwalających na uczenie i implementację sieci różnej precyzji na platformach MPSoC firmy Xilinx AMD. Sieć uczona z narzędziem Brevitas osiągnęła niższą skuteczność (95%), natomiast z odpowiednią konfiguracją akceleratora generowanego przez narzędzie FINN zaproponowano detektor przetwarzający ponad 580 FPS przy poborze energii 3.547W. Szczegóły dotyczące implementacji sprzętowych systemów detekcji znaków drogowych przedstawiono w [Przewłocka and Kryjak, 2019] oraz [Przewłocka-Rus et al., 2021].

Korzystając z tych samych narzędzi przeprowadzono szereg badań związanych z akceleracją systemów śledzenia obiektów, w szczególności rozwiązania wyznaczającego w momencie rozpoczęcia badań SOTA SiamFC, bazującego na Syjamskich Sieciach Neuronowych. W ramach eksperymentów pokazano, że w przypadku architektur nadmiarowych (tzw. *dużych* sieci neuronowych) kwantyzacja liniowa warstw ukrytych może wpłynąć pozytywnie na skuteczność działania modelu. Ponadto zwrócono szczególną uwagę na wyzwania związane z próbami bezpośredniej implementacji rozwiązań SOTA w urządzeniach FPGA/SoC i konieczność projektowania algorytmów w sposób dopasowany do docelowych wbudowanych platform wykonawczych (tzw. podejście *hardware aware algorithm co-design*). Zaproponowano alternatywną architekturę sieci syjamskiej redukującą liczbę parametrów prawie 7-krotnie względem rozwiązania wzorcowego SOTA, jednocześnie gwarantując wysoką skuteczność śledzenia. Dalszej redukcji dokonano kwantyzując wagi warstw ukrytych do liczb całkowitych 4-bitowych, i warstw

pierwszej i ostatniej do liczb całkowitych 8-bitowych, korzystając z kwantyzacji liniowej. Tym samym zaproponowano niewielką w sensie złożoności pamięciowej architekturę, która pozwalała na działanie systemu śledzącego ze skutecznością zbliżoną do modelu pełnej precyzji (ze spadkiem mniejszym niż 2%). Następnie sieć uruchomiono na platformie Zynq UltraScale+ MPSoC ZCU104 osiągając prawie 50 FPS przy poborze energii 5.5 W. Pełny algorytm śledzenia, uproszczony do przetwarzania ROI w jednej skali, zaimplementowany częściowo również w części procesorowej układu Zynq, osiąga 17 FPS pobierając 5.5 W, w porównaniu do oryginalnego systemu uruchamianego w NVIDIA GeForce GTX Titan X z poborem mocy 250 W (ok. 45 razy więcej) i osiągającym 83 FPS. Szczegóły metod i implementacji sprzętowych opisano w [Przewłocka et al., 2020], [Przewłocka-Rus and Kryjak, 2021] oraz [Przewłocka-Rus and Kryjak, 2022b].

Podsumowanie

W ramach podsumowania prowadzonych badań i zaproponowanych metod, warto zwrócić uwagę na dwie istotne kwestie. Po pierwsze, jak pokazano przy okazji prac nad logarytmiczną kwantyzacją PoT, specjalne schematy kwantyzacji, dostosowane do dystrybucji wag w warstwach sieci neuronowych mogą nie tylko gwarantować znacznie niższy spadek skuteczności względem modelu pełnej precyzji, ale również pozwalają na znaczne uproszczenia architektury obliczeniowej akceleratorów sieci neuronowych. Po drugie, pokazano też w jaki sposób akcelerować zaawansowane systemy wizyjne w urządzeniach FPGA/SoC, proponując odpowiednie modyfikacje rozwiązań SOTA w taki sposób, by były one możliwe do zrealizowania sprzętowo, również wykorzystując istniejące narzędzia.

Do najważniejszych oryginalnych osiągnięć autorki należą:

1. Zaproponowanie metod uczenia sieci kwantyzowanych do wag o wartościach potęg dwójki, pozwalających osiągnąć skuteczność na poziomie modeli pełnej precyzji nawet dla niskiej szerokości bitowej (4 bity).

2. Zaprojektowanie architektury sprzętowej realizującej operację MAC dla sieci PoT z użyciem operacji przesunięcia bitowego zamiast mnożenia. Zaproponowany pojedynczy element obliczeniowy dla kwantyzacji 4x8 (szerokość wag x szerokość aktywacji) pozwala na redukcję zapotrzebowania energetycznego x2 względem modułu dla modelu kwantyzowanego liniowo z takimi samymi szerokościami bitowymi.

3. Zaprojektowanie architektury sprzętowej warstwy konwolucyjnej PoT, która używa ok. 0.6 energii potrzebnej dla standardowej warstwy i zwiększa zakres częstotliwości pracy układu.

4. Zaproponowanie metody fuzji warstw konwolucyjnej i normalizacji partiami, uwzględniającej szczególną postać sieci PoT.

5. Przeprowadzenie szeregu eksperymentów i analizy wpływu różnych schematów kwantyzacji (liniowej, logarytmicznej), z różną docelową szerokością bitową, dla sieci neuronowych używanych w zaawansowanych systemach wizyjnych, celem oceny wpływu na skuteczność, złożoność pamięciowo-obliczeniową oraz energooszczędność takich systemów.

6. Zaprojektowanie algorytmiczno-sprzętowych zaawansowanych i energooszczędnych systemów wizyjnych działających z użyciem głębokich sieci neuronowych: detekcji znaków drogowych, śledzenia obiektów oraz detekcji pieszych i pojazdów.

Zaproponowane metody kwantyzacji i ich odpowiednie użycie do redukcji złożoności pamięciowo-obliczeniowej wybranych modeli sieci neuronowych, wraz z odpowiednią orkiestracją obliczeń, umożliwiają implementację systemów wizyjnych czasu rzeczywistego w urządzeniach niewielkiej mocy, co dowodzi słuszności

postawionej tezy. Wyniki opisanych wyżej badań opublikowano w serii 9 artykułów zestawionych w tabeli 3, wraz z liczbą cytowań (w sumie: 57 (51), stan na dzień 24 lutego 2025, w nawiasie podano liczbę bez autocytowań).

Tabela 3: Skrótowa lista publikacji wchodzących w cykl publikacji, z liczbą cytowań (stan na dzień 24 lutego 2025), w nawiasach podano liczbę bez autocytowań)

Publikacja	**Tytuł**	**Cytowania**
[Przewłocka and Kryjak, 2019]	*XNOR CNNs in FPGA: real-time detection and classification of traffic signs in 4K – a demo*	1 (1)
[Przewłocka et al., 2020]	*Optimisation of a Siamese neural network for real-time energy efficient object tracking*	6 (5)
[Przewłocka-Rus and Kryjak, 2021]	*Quantised Siamese tracker for 4K/UltraHD video stream – a demo*	0 (0)
[Przewłocka-Rus et al., 2021]	*Exploration of hardware acceleration methods for an XNOR traffic signs classifier*	0 (0)
[Przewłocka-Rus et al., 2022]	*Power-of-two quantization for low bitwidth and hardware compliant neural networks*	37 (34)
[Przewłocka-Rus and Kryjak, 2022b]	*Towards real-time and energy efficient Siamese tracking – a hardware-software approach*	5 (5)
[Przewłocka-Rus and Kryjak, 2022a]	*Energy efficient hardware acceleration of neural networks with power-of-two quantisation*	6 (4)
[Przewłocka-Rus and Kryjak, 2023]	*Power-of-Two Quantized YOLO Network for Pedestrian Detection with Dynamic Vision Sensor*	2 (2)
[Przewłocka-Rus et al., 2024]	*PowerYOLO: Mixed Precision Model for Hardware Efficient Automotive Detection with Event Data*	0 (0)

Bibliografia

[Przewłocka and Kryjak, 2019] Przewłocka, D. and Kryjak, T. (2019). XNOR CNNs in FPGA: real-time detection and classification of traffic signs in 4K – a demo. *DASIP 2019: Conference on Design and Architectures for Signal and Image Processing*. 16–18 October 2019, Montréal, Canada.

[Przewłocka et al., 2020] Przewłocka, D., Wąsala, M., Szolc, H., Błachut, K., and Kryjak, T. (2020). Optimisation of a Siamese neural network for real-time energy efficient object tracking. *Chmielewski, L.J., Kozera, R., Orłowski, A. (eds) Computer Vision and Graphics. ICCVG 2020. Lecture Notes in Computer Science, vol 12334. Springer, Cham.*

[Przewłocka-Rus et al., 2021] Przewłocka-Rus, D., Kowalczyk, M., and Kryjak, T. (2021). Exploration of hardware acceleration methods for an XNOR traffic signs classifier. *Choraś, M., Choraś, R.S., Kurzyński, M., Trajdos, P., Pejaś, J., Hyla, T. (eds) Progress in Image Processing, Pattern Recognition and Communication Systems. CORES IP&C ACS 2021. Lecture Notes in Networks and Systems, vol 255. Springer*.

[Przewłocka-Rus and Kryjak, 2021] Przewłocka-Rus, D. and Kryjak, T. (2021). Quantised Siamese tracker for 4K/UltraHD video stream – a demo. *2021 31st International Conference on Field-Programmable Logic and Applications (FPL)*. 30 August - 3 September 2021, Dresden, Germany.

[Przewłocka-Rus and Kryjak, 2022a] Przewłocka-Rus, D. and Kryjak, T. (2022a). Energy efficient hardware acceleration of neural networks with power-of-two quantisation. *Chmielewski, L.J., Orłowski, A. (eds) Computer Vision and Graphics. ICCVG 2022. Lecture Notes in Networks and Systems, vol 598. Springer, Cham.*

[Przewłocka-Rus and Kryjak, 2022b] Przewłocka-Rus, D. and Kryjak, T. (2022b). Towards real-time and energy efficient Siamese tracking – a hardware-software approach. *Desnos, K., Pertuz, S. (eds) Design and Architecture for Signal and Image Processing. DASIP 2022. Lecture Notes in Computer Science, vol 13425. Springer, Cham.*

[Przewłocka-Rus and Kryjak, 2023] Przewłocka-Rus, D. and Kryjak, T. (2023). Power-of-Two Quantized YOLO Network for Pedestrian Detection with Dynamic Vision Sensor. *26th Euromicro Conference on Digital System Design (DSD)*. 6-8 September 2023, Durres, Albania.

[Przewłocka-Rus et al., 2024] Przewłocka-Rus, D., Kryjak, T., and Gorgon, M. (2024). PowerYOLO: Mixed Precision Model for Hardware Efficient Automotive Detection with Event Data. *27th Euromicro Conference on Digital System Design (DSD)*. 28-30 August 2024, Sorbonne University, Paris France.

[Przewłocka-Rus et al., 2022] Przewłocka-Rus, D., Sarwar, S. S., Sumbul, H. E., Li, Y., and Salvo, B. D. (2022). Power-of-two quantization for low bitwidth and hardware compliant neural networks. *TinyML Research Symposium 2022*. 28 march 2022, San Jose, USA. Available at:`https://cms.tinyml.org/wp-content/uploads/talks2022/2203.05025.pdf`.