Prof. dr hab. inż. Jerzy Stefanowski,
Instytut Informatyki Politechnika Poznańska
Polska Akademia Nauk.

Poznań, 27.05. 2026

Recenzja rozprawy doktorskiej mgra inż. Bartosza Wójcika pt. „Computationally Efficient Neural Networks"

1. Uwagi ogólne nt. problematyki rozprawy i jej celów

Niniejsza recenzja jest sporządzona w odpowiedzi na pismo wiceprzewodniczącego Rady Dyscypliny Informatyka Techniczna i Telekomunikacja Uniwersytetu Jagiellońskiego.

Rozprawa dotyczy problematyki uczenia głębokich sieci neuronowych w zakresie nowej propozycji redukcji sieci prowadzących do mniejszych kosztów obliczeniowych przy zachowaniu trafności predykcji porównywalnej z oryginalnymi sieciami. Motywacje do takiego działania są dość szeroko przedstawione we wstępie do maszynopisu i są uzasadnione w kontekście obecnego „kosztownego" uczenia i stosowania głębokich sieci neuronowych.

Pomimo, że badania nad różnymi podejściami do przycinania (ang. pruning) nawet prostszych sieci wielowarstwowych były prowadzone od dziesiątków lat, później rozważano kwantyzacje, „destylacje" elementów architektury sieci i inne rozwiązania specyficzne dla współczesnego uczenia głębokiego jest to nadal otwarta i nowoczesna problematyka badawcza.

W swoich badaniach Pan Bartosz Wójcik skupił uwagę najpierw na optymalizacji głębokości sieci, a później na dynamicznej redukcji elementów w szerokości architektury sieci. Większość z rozważanych przez doktoranta sieci jest specyficzna dla przetwarzania i klasyfikacji obrazów, a niektóre późniejsze sieci są też dostosowane do przetwarzania tekstów.

W mojej opinii przedstawione przez Doktoranta motywacje krytyczne oraz sformułowane cele badawcze są nowatorskie i warte podjęcia badań w rozprawie doktorskiej.

2. Ocena formy i struktury rozprawy

Przekazany maszynopis rozprawy jest napisany w języku angielskim i obejmuje kopie pięciu następujących artykułów – w większości prezentacji konferencyjnych (które stanowią osiągnięcie Doktorantki):

1. Wójcik B., Wołczyk M., Bałazy K., Tabor J. Finding the optimal network depth in classification tasks. Proc. ECMLPKDD 2020.
2. Wołczyk M., Wójcik B., Bałazy K., Podolak I., Tabor J. Śmieja M., Trzciński T. Zero time waste: Recycling predictions in early exit neural networks, Proc. NeurIPS 2021.
3. Wójcik B., Przeźlikowski M., Szatkowski F., Wołczyk M. Bałazy K., Krzepkowski B., Podolak I., Tabor J. Śmieja M., Trzciński T. Zero time waste in pre-trained nearly exit neural networks, Neural Networks Journal 2023.
4. Wójcik B., Devoto A., Pustelinik K., Minervinini P. Scardapane S. Adaptive computational modules: Granular conditional computation for efficient inference. Proc. AAAI 2025.
5. Szatkowski F., Wójcik B., Piórczynski M., Scardapane S. Exploiting activation sparsity with dense to dynamic k mixture of experts conversion. Proc. NeurIPS 2024.

Powyższe prace były przedstawione i opublikowane na bardzo znanych konferencjach z zakresu uczenia maszynowego oraz sztucznej inteligencji. W tym konferencja NeurIPS jest obecnie bardzo trudna do pozytywnego „przejścia" poprzez dość specyficzny proces recenzyjny. Pozycja 3 jest artykułem w dość dobrym czasopiśmie z zakresu sieci neuronowych.

Wszystkie artykuły są mocno wieloautorskie (od czterech współautorów do dziesięciu – co zastanawia, dlaczego aż tylu) = przy czym Pan Bartosz Wójcik nie jest zawsze umieszczony na pierwszym miejscu w niealfabetycznych listach autorów. Na stronie 11 Doktorant pisze „I am the primary author of each work in the cycle" – co nie jest jednak udokumentowane w inny sposób – zwłaszcza, że przy niektórych kopiach tekstów jest podany równy wkład z innymi autorami / więc może to jednak budzić wątpliwości. W opisie poszczególnych prac Pan Bartosz Wójcik umieszcza krótki akapit „contribution", gdzie precyzuje swój wkład do pracy poprzez wymienienie tzw. aktywności [nie podaje się wkładu procentowego, co zwłaszcza dla prac 2 i 3 byłoby ciekawe] – w mojej opinii nie jest on równy pomiędzy pracami. W ramach uwag krytycznych uważam, że przy tak wielo-autorskich pracach Doktorant mógł precyzyjniej opisać swoją rolę jako „pierwszego autora".

Odnosząc się do wymagań ustawowych rozprawa doktorska nie musi być typowym maszynopisem, lecz może być przedstawiona w formie powiązanego tematycznie zbioru artykułów (opublikowanych lub przyjętych do druku w czasopismach naukowych, określonych przez Ministra właściwego do spraw nauki) – art. 187. Ponadto inny artykuł ustawy stwierdza, że Doktorant powinien mieć choć jeden artykuł opublikowany w czasopiśmie naukowym lub w recenzowanych materiałach z konferencji międzynarodowej, które w roku opublikowania artykułu w ostatecznej formie były ujęte w wykazie sporządzonym przez Ministra – co jest w przypadku recenzowanej rozprawy jest spełnione dla wszystkich z przedstawionych pięciu prac.

Także w mojej opinii spełniono wymóg powiązania tematycznego – wszystkie prace mieszczą się w tematyce rozprawy i są zgodne z przedstawionym celem badań doktorskich. Samo ułożenie kolejności tych prac jest poprawne i przedstawia stopniową ewolucję kolejnych metod redukcji sieci.

Sam maszynopis rozprawy, oprócz kopii zestawu pięciu artykułów, zawiera:

- Bardzo krótkie wprowadzenie do tematyki rozprawy i badan (str. 7-12) w ramach którego zdefiniowano cel rozprawy – str. 9 – „zredukować koszty obliczeniowe głębokich sieci neuronowy zachowując ich zdolności predykcyjne", oraz przedstawiono wstępnie osiągnięcia każdej z prac cyklu.
- Dwa rozdziały opisujące – główne osiągnięcia każdej prac, zalety oraz ograniczenia oraz wkład własny autora. Rozdział 2gi obejmuje trzy prace, w których redukowano głębokość struktury statycznej sieci (str. 12 -26). Rozdział 3 przedstawia podejścia dynamiczne do przebudowania jednostek w „szerokości" topologii sieci (str. 28-37).
- Streszczenia polsko- i anglojęzyczne (str. 3-4)
- Opis aktywności naukowych i zawodowych Doktoranta – w tym listę pięciu dodatkowych artykułów oraz udział w realizacji projektów badawczych (w tym projektu NCN Preludium)– co oceniam bardzo pozytywnie – w mojej opinii liczba ciekawych prac jest wyższa niż standardowo oczekujemy od doktorantów.
- Listę cytowanych publikacji w rozdziałach maszynopisu.

Po tym elementach maszynopis zawiera kopie pięciu artykułów stanowiących osiągnięcie Doktoranta. Pewnym brakiem – jest niedodanie do wszystkich artykułów tzw. załączników, o których piszę się w samym maszynopisie albo treści artykułów.

Z punktu widzenia czytelnika jest to dość dobrze i logicznie skonstruowany maszynopis. Język angielski maszynopisu jest poprawny i dostatecznie jasny.

3. Ocena dokonanie Doktoranta

Po zapoznaniu się z maszynopisem, w szczególności przeczytaniu wskazanych artykułów = osiągnięcia Doktoranta i jego współautorów w zakresie redukcji obliczeń w głębokich sieciach neuronowych można podzielić na dwa kierunki badawcze:

1. Możliwości redukcji liczby warstw w sieci, tj. jej głębokości, poprzez trzy różne metody wywodzące się z inspiracji tzw. metody wczesnych wyjść oraz jej uogólnień w stronę specyficznego zespołu predykcyjnego (artkuły 1 – 3).
2. Podejść do zastępowania części neuronów (podsieci) w szerokości warstw sieci przez nowe jednostek obliczeniowe, które odpowiadają koncepcji złożenia eksperckiego (odpowiednik ang. mixture of experts) i mogą dynamicznie dostosować obliczenia dla indywidualnych (trudnych albo łatwiejszych) przykładów podczas wnioskowania (artykuły 4-5).

Te kierunki działań są w mojej opinii trochę ortogonalne (np. pierwsze prace są bardziej statyczne i dotyczą się dostosowania całej struktury do całego zbioru danych), podczas gdy większość propozycji ma charakter dynamiczny dostosowujący się do trudności i specyfiki aktualnie przetwarzanego przykładu.
Ocena propozycji jest zawsze wykonywana eksperymentalnie, gdzie autorzy prac starają się pokazywać, że zmniejszona sieć wystarczająco zachowuje trafność predykcji, a w części prac (raczej późniejszych) pokazywali także wyraźniej przetarg z narzutami obliczeniowymi. Większość eksperymentów wykonywano na znanych i popularnych danych obrazowych oraz późniejsze dynamiczne metody złożeń ekspertów także dla danych tekstowych w języku naturalnym. Nie w każdym artykule wykorzystywano liczniejsze kolekcje zbiorów danych (np. w pierwszej pracy tylko dwa) co jest pewnym ograniczeniem. Z kolejnymi pracami zwiększała się liczba porównywanych metod do redukcji rozmiarów sieci (co oceniam pozytywnie) oraz prowadzono obszerniejsze studia wpływu części składowych proponowanych rozwiązań na wyniki końcowe (tzw. ang. ablation studies). Trudno oceniać strukturę prac, które zostały już zrecenzowane i opublikowane – lecz wydaje mi się, że są napisane dość typowo dla obecnie publikowanych na tych konferencjach i są w miarę poprawne. Ponadto chcę pozytywnie ocenić, że w każdej z prac są linki do repozytoriów z kodem – nie rozwiązuje to w pełni problemu odtwarzalności (wymagałoby to udostępniania szerszego zakresu dodatkowych danych), lecz jest to dobre zachowanie ze strony autorów. Poniżej opisuję i oceniam najważniejsze z przedstawionych osiągnięć.

Pierwsza propozycja metody redukcji NETCUT (artykuł 1) jest dla mnie modyfikacją wcześniejszych propozycji metod tzw. wcześniejszych wyjść z kolejnych warstw sieci (ang. early exits), gdzie zaproponowano nową funkcje łącząca głowice z tych wyjść warstw co ma ułatwić w optymalizacji wybór jednej z głowic (wyjść sieci), czyli wskazanie wcześniejszej warstwy. Propozycja jest oceniona eksperymentalnie z typowymi obrazowymi sieciami konwolucyjnymi na dwóch popularnych zbiorach danych obrazowych oraz, co jest ciekawe, dla wersji sieci grafowej pokazując, że można zredukować liczbę warstw sieci nadal oferując podobną trafność co cała sieć.

Dwa kolejne artykuły dotyczą w zasadzie tej samej propozycji Zero Waste / Time Early Exiting (w skrócie ZTW) w innej fazie działania sieci, to jest wnioskowania dla nowego przykładu. Intuicyjnie propozycja jest modyfikacją poprzednio omawianych wczesnych wyjść, gdzie wprowadzono sygnał z wcześniejszych jednostek wyjść do kolejnych, co przy odpowiednio zaproponowanym podejściu do warunków sprawdzanych na wyjściu może doprowadzić do „szybszych" i dostatecznie pewnych predykcji. Jak sami autorzy piszą można popatrzeć na ten pomysł jak pośrednio inspirowany konstrukcją zespołów predykcyjnych. Pierwsza z praca (artykuł 2), pomimo publikacji jako poster na NeurIPS, wydaje mi się tylko wstępnym pokazaniem pomysłów z dość (dla mnie) mocnym ograniczeniem eksperymentów, np. tylko dwie w miarę łatwe do pobicia rozwiązania konkurencyjnych sieci. Za ważniejszy uważam artykuł nr 3, który mogę ocenić za lepsze rozszerzenie poprzednio przedstawionej koncepcji poprzez mocniej rozbudowaną analizę eksperymentalną (z bardziej konkurencyjnymi metodami), dokładniejszą analizą roli poszczególnych składników ZTM, np. badaniu czy propozycja soft-max oszacowania prawdopodobieństw jest dość dobra z punktu widzenia kalibracji wyjść jednostek obliczeniowych oraz możliwości integracji z technikami

destylacji wiedzy z sieci oraz tzw. transfer learning pomiędzy różnymi danymi (wersje zbiorów CIFAR). Pomimo, że wyniki eksperymentalne artykułu 3 są bardziej zachęcające niż wstępne z pracy 2, to jak sam doktorant zauważa w maszynopisie na str. 26 maszynopisu te propozycje mają umiejscowienie tylko w działaniach na głębokości sieci głównie poprzez ocenę możliwości wcześniejszego podjęcia decyzji o wykonaniu predykcji.

Rozszerzając powyższe motywacje, Doktorat słusznie zauważył, że redundancja struktury sieci może występować także wewnątrz samych warstw i możliwe byłoby też dynamicznie określanie, które z nich powinny brać udział w obliczeniach dla konkretnego przetwarzanego przykładu. Wiąże się to z inną obserwacją, że niektóre przykłady (zwłaszcza obrazy) nie potrzebują wykorzystywania wszystkich neuronów obliczeniowych dla ich łatwiejszej części – więc można dynamicznie pod trudność przykładu zarządzać przydziałem liczby jednostek obliczeniowych. Prowadzi to doktoranta i współautorów do nowej koncepcji obniżania obliczeń. W pracy (artykuł nr 4) zaproponowali zastępowanie fragmentów sieci specjalnymi blokami jednostek obliczeniowych, które mogą być składane i później pojedyncze wybierane poprzez wyuczoną funkcję bramkującą = to rozwiązane jest silnie inspirowane poprzednimi pracami na koncepcją złożenia sieci ekspertowych - tzw. ang. mixture of experts. Wydaje mi się, że oryginalnym pomysłem jest też strategia inicjalizacji parametrów modeli ACM, opartą na destylacji wiedzy na poziomie modułów z wcześniej wytrenowanych modeli. Pomimo inspiracji można dostrzec oryginalne pomysły, które pozwalając na ciekawe rozwiązanie, które z pewną granularnością pozwala lepiej analizować i przydzielać zasoby do podregionów obrazów oraz zapisów mowy – co pokazano w eksperymentach. Ciekawe są także wyniki eksperymentalne frontów Pareto (Accuracy, czas wnioskowania), gdzie widać dominacje proponowanego rozwiązanie ACM.

Trochę inne rozwinięcie tego rodzaju podejścia jest widoczne z ostatniej pracy cyklu – nr 5. Jest ona dodatkowo inspirowana obserwacja, że przy przetwarzaniu obrazów lub tekstów wiele z neuronów w warstwach (zwłaszcza przy stosowaniu funkcji przejścia ReLU) jest nieaktywne – co można wykorzystać do dalszej redukcji. Prowadzi to autorów i doktoranta do zaproponowania dość złożonego podejścia o nazwie D2DMoE, które rozszerza wcześniejsze prace nad Mixture of experts o takie elementy jak ciekawe dostrojenie wstępnie wytrenowanej sieci w celu zapewnienia rzadkości wartości aktywacji wielu neuronów w warstwie. Następnie grupuje się neurony w warstwach sieci, tworząc odrębne modele eksperckie i wprowadza element podsieci routującej. Na koniec sieć routerowa jest precyzyjnie dostrajana w celu przewidywania odpowiedniej normy na wyjściu każdego eksperta. Dodatkowo w niniejszej pracy badana jest możliwość modyfikacji warstw uwagi oraz grupowania warstw MLP przy użyciu funkcji aktywacji GLU. Wyniki eksperymentów (głownie na danych językowych i sieciami typu prostsze transformery) pokazują, że proponowana metoda jest odporna na wysoki stopień rzadkości i przy tworzeniu frontu Pareto przewyższa kilka porównywanych wcześniejszych metod – w tym własną ZTW.

Podsumowując tą część recenzji zdecydowanie uważam, że powyższe prace zawierają interesujące propozycje rozwiązanie celu rozprawy. Choć z punktu widzenia inspiracji można też stwierdzić, że niektóre z nich mają mniejszy stopień oryginalności, gdyż rozwijają wcześniejsze propozycje z literatury albo łączą inspiracje z kilku różnych podejść, to widać własne dodatkowe propozycje w zakresie definiowania funkcji straty, oszacowań prawdopodobieństw wyjść z podsieci, itd. Ponadto ich ocena eksperymentalna jest bardzo interesująca i zachęcająca do dalszych badań. Dla mnie najbardziej oryginalne i potencjalnie interesujące do dalszych badań są propozycje rozwijającą podejścia złożenia ekspertów z dwóch ostatnich artykułów. Ponadto uważam, że Pana Bartosz Wójcik starał się rozwiązać dużo więcej problemów badawczych niż oczekujemy wobec standardowych rozpraw doktorskich.

4. Uwagi krytyczne i polemiczne

Pomimo interesujących propozycji rozwiązywania ważnego problemu i mojej akceptacji dla zaproponowania wielu wersji możliwych sposobów redukcji struktury sieci i obniżania jej kosztów obliczeniowych formułuję kilka uwag krytycznych, co do treści niektórych artykułów oraz pytania o charakterze dyskusji polemicznej z Doktorantem, które mogłyby być przedmiotem dalszej dyskusji podczas samej obrony doktoratu. Pisząc to jestem też świadomy, że prace z cyklu przeszły dość rygorystyczne recenzowanie w części konferencji, więc może wiele elementów i tak uległo w ostatecznej wersji tekstów poprawie.

Po pierwsze jako badacz przyzwyczajony do trochę innych systemów uczących się (może bardziej tradycyjnych niż rozważane przez Doktoranta i współautorów) czytając niektóre z prac miałem wrażenie, że niektóre eksperymenty są dość ograniczone w liczbie porównywanych metod. Na przykład w ostatniej z prac D2DMoE jest główne porównania praktycznie z prostszym MoEfication (z którym wygrana jest chyba spodziewana) i ZTW = który zaproponowali autorzy do innych celów redukcji. Podobnie mógłbym skomentować prace 1 i 2, gdzie w tzw. „related works" jest opisane więcej konkurencyjnych prac. Można też podobne obserwacje sformułować do wykorzystywania też nie za dużo benchmarków obrazowych (raczej b. popularne i znowu tylko kilka na artykuł). Przykładowo w pracy 2 miałem wrażenie, eksperyment porównawczy nie odnosi się do wielu metod mogących być tzw. SOTA i danych – co czytając prace 3 (która ma szersze eksperymenty) potwierdziło się. Prosiłbym Doktoranta o komentarz nt. zasad wyboru porównywanych metod – przynajmniej w pracach 1,2 i przede wszystkim 5tej.

Większość eksperymentów jest ukierunkowana na pokazanie wyników utrzymania trafności predykcji przez prostsze sieci – w mniejszym stopniu pokazuje się wyniki dla czasów obliczeń i innych aspektów kosztów obliczeń. Czy mógłbym prosić o komentarz nt. metodologii takiej oceny.

W późniejszych pracach stosuje się wykresy dwukryterialne (Accuracy vs. FLOPS) – które są oczywiście korzystne dla autorów prac. Lecz trochę mam wątpliwości czy miara FLOPS jest wystarczająca, i czy nie warto było poszukiwać także innych mierników kosztów obliczeń. Ponadto czy aspekt kosztów nie będzie powiązany z implementacjami i obliczeniami na konkretnych kartach GPU – i nie powinno być to odpowiednio uwzględniane w dyskusji.

Powiązanym pytaniem jest czy niektóre rozwiązanie – np. tworzenie zestawów sieci ekspertów, odpowiednich funkcji przełączających oraz routera (w ostatnich pracach) albo podsieci do wczesnych wyjść nie powodują dodatkowych narzutów, także pamięciowych. Nie znalazłem tych wątków, aby były dyskutowana przez Doktoranta – prosiłbym o wyjaśniania podczas obrony.

Przyglądając się niektórym wynikom samej trafności sieci na danych obrazowych w tabelach (np. ZTM Tabela 2.1 rozdział 2.2, albo Tabela 1 artykuł 2) różnice pomiędzy porównywanymi algorytmami zwłaszcza SDN vs. proponowane ZTM są dość małe, nawet w przypadku (Tabela 1 artykuł 2 – sieci jak MobileNet i VGG) praktycznie ich nie ma. Proszę o komentarz, dlaczego w takich sytuacjach twierdzono, że ZTM jest lepsza i czy te różnice są statystycznie znaczące w konsekwencji określonych scenariuszy testowania modeli.

W przypadku niektórych propozycji widać podobieństwo rozwiązań do innych z opisywanych w literaturze – wydaje mi się, że o ile w opublikowanych artykułach trudno było to już rozwijać, to w samym maszynopisie w ramach dodatkowych rozdziałów Doktorant powinien był to lepiej rozważyć, aby mocniej dyskutować tzw. nowatorskość propozycji. Dla przykładu proszę o rozszerzający komentarz nt. podobieństw pomiędzy własną metodą ACM (artykuł 4) a wzmiankowaną propozycją Cai et al. (pozycja nr 8 w literaturze maszynopisu).

Ostatnia uwaga – zarówno w rozdziałach jak i niektórych pracach piszę się, że dodatkowe obliczenia wykonano w załącznikach (patrz np. praca 1 o NetCut) lecz nie są one dodane do konkretnych prac w maszynopisie.

5. Konkluzja końcowa

Podsumowując uważam, że przedstawiona rozprawa charakteryzuje się dobrym poziomem merytorycznym i zawiera interesujące rezultaty badawcze, które mogą być przydatne dla rozwoju głębokich sieci neuronowych. Założone cele badań doktorskich zostały osiągnięte.

Dlatego uważam, że rozprawa Pana mgra inż. Bartosza Wójcika spełnia warunki stawiane przez ustawę prawo o szkolnictwie wyższym i nauce w odniesieniu do rozpraw doktorskich i może być dopuszczona do publicznej obrony.

Jerzy Stefanowski

Prof. Jerzy Stefanowski