Warszawa, 05.03.2026

UNIWERSYTET
WARSZAWSKI
Wydział Matematyki, Informatyki i Mechaniki
Instytut Informatyki

Recenzja rozprawy doktorskiej magistra inżyniera Bartosza Wójcika pod tytułem *Computationally Efficient Neural Networks*

Artykuły na których bazuje rozprawa

[P1] B. Wójcik, M. Wołczyk, K. Bałazy, J. Tabor. *Finding the optimal network depth in classification tasks*. ECML PKDD 2020.

[P2] M. Wołczyk, B. Wójcik, K. Bałazy, I. Podolak, J. Tabor, M. Śmieja, T. Trzciński. *Zero Time Waste: Recycling Predictions in Early Exit Neural Networks*. NeurIPS 2021.

[P3] B. Wójcik i współautorzy. *Zero Time Waste in Pre-trained Early Exit Neural Networks*. Neural Networks, 2023.

[P4] B. Wójcik i współautorzy. *Adaptive Computation Modules: Granular Conditional Computation for Efficient Inference*. AAAI 2025.

[P5] F. Szatkowski, B. Wójcik i współautorzy. *Exploiting Activation Sparsity with Dense to Dynamic-k Mixture-of-Experts Conversion*. NeurIPS 2024.

Tematem rozprawy są metody zwiększania efektywności obliczeniowej modeli głębokiego uczenia poprzez wykorzystanie mechanizmów dynamicznej alokacji obliczeń. W ostatniej dekadzie obserwujemy bardzo dynamiczny rozwój metod uczenia głębokiego, który w dużej mierze został umożliwiony przez rosnącą moc obliczeniową współczesnych procesorów graficznych (GPU). Jednocześnie coraz większe modele prowadzą do znaczącego wzrostu kosztów treningu i wnioskowania, zarówno finansowych jak i energetycznych. W związku z tym

rozwój metod umożliwiających zmniejszenie kosztów obliczeniowych przy zachowaniu wysokiej jakości predykcji stanowi istotny i aktualny kierunek badań.

Rozprawa składa się z cyklu pięciu publikacji naukowych poświęconych różnym aspektom redukcji redundancji obliczeniowej w modelach głębokiego uczenia. Prace te można podzielić na dwie główne części: metody redukcji kosztów obliczeniowych związanych z głębokością sieci oraz metody związane z dynamiczną adaptacją szerokości modeli.

Zanim przejdę do omówienia wyników przedstawionych w pracy chciałbym podkreślić bardzo przejrzystą strukturę rozprawy. Autor jasno identyfikuje źródła redundancji obliczeniowej w modelach głębokiego uczenia oraz systematycznie analizuje różne podejścia do ich eliminacji. Takie uporządkowanie materiału znacząco ułatwia zrozumienie motywacji stojącej za kolejnymi publikacjami. Doceniam wysiłek autora włożony w przygotowanie omówienia prac wchodzących w skład rozprawy, gdyż omówienie to zawiera spójną i obszerną narrację omawiającą kolejne wyniki i ich zależności.

W pracy [P1] autor zajmuje się problemem wyboru optymalnej głębokości sieci neuronowej. Autor proponuje podejście polegające na dodaniu klasyfikatorów pomocniczych do kolejnych warstw sieci i uczeniu wag określających znaczenie poszczególnych poziomów reprezentacji. W wyniku optymalizacji model sam wybiera najbardziej użyteczną głębokość, co pozwala na usunięcie zbędnych warstw po zakończeniu treningu.

Prace [P2] oraz [P3] dotyczą metod typu early-exit, które umożliwiają dynamiczne zakończenie obliczeń w trakcie propagacji przez sieć. Intuicja stojąca za tym podejściem jest taka, że dla wielu przykładów wejściowych pełna głębokość modelu nie jest konieczna do uzyskania poprawnej klasyfikacji. W pracy [P2] zaproponowana została metoda Zero Time Waste, która polega na ponownym wykorzystaniu predykcji generowanych przez wcześniejsze klasyfikatory pomocnicze. Dzięki temu możliwe jest ograniczenie zjawiska określanego w literaturze jako overthinking, gdzie kolejne warstwy modelu zmieniają poprawną predykcję na błędną. W pracy [P3] podejście to zostało rozszerzone o szeroką analizę eksperymentalną obejmującą m.in. scenariusze transfer learning oraz analizę kalibracji predykcji.

Kolejne dwie prace [P4] i [P5] dotyczą problemu redukcji kosztów obliczeniowych związanych z szerokością modeli. W pracy [P4] autor proponuje moduł Adaptive Computation Module, który umożliwia dynamiczne dostosowanie liczby wykonywanych operacji do trudności poszczególnych elementów danych wejściowych. W szczególności obliczenia mogą być wykonywane z różną intensywnością dla różnych fragmentów obrazu lub sekwencji. Jest to interesujące rozwinięcie idei conditional computation na bardziej granularnym poziomie niż klasyczne metody early-exit.

Praca [P5] dotyczy z kolei metod opartych o architekturę Mixture-of-Experts. Autor analizuje związek pomiędzy rzadkością aktywacji neuronów a możliwością dynamicznego wyboru ekspertów w trakcie wnioskowania. Zaproponowana metoda Dense-to-Dynamic MoE

umożliwia konwersję modeli gęstych do architektury dynamicznej, w której liczba aktywowanych ekspertów zależy od danych wejściowych. W pracy pokazano również praktyczne aspekty implementacyjne związane z wydajnością takich modeli na współczesnych architekturach GPU.

Warto również zwrócić uwagę, że publikacje składające się na rozprawę powstawały na przestrzeni około sześciu lat. W tym czasie dziedzina uczenia maszynowego rozwijała się niezwykle dynamicznie, a wiele kierunków badań związanych z efektywnością obliczeniową modeli ulegało istotnym zmianom. Pojawiały się nowe architektury oraz nowe paradygmaty trenowania modeli, które częściowo zmieniły perspektywę patrzenia na problem kosztów obliczeniowych. Z tego względu interesujące byłoby, gdyby w trakcie obrony autor odniósł się do swoich wyników z szerszej perspektywy czasu i spróbował ocenić rozwój całej dziedziny, nie tylko własnego dorobku. W szczególności ciekawa byłaby refleksja nad tym, które z proponowanych w ostatnich latach podejść do zwiększania efektywności modeli – takich jak early-exit, conditional computation czy architektury typu mixture-of-experts – okazały się trwałym elementem kanonu metod uczenia maszynowego, a które pozostały rozwiązaniami bardziej niszowymi. Tego rodzaju syntetyczna perspektywa mogłaby stanowić wartościowe uzupełnienie przedstawionych w rozprawie wyników.

Ocena rozprawy

Podsumowując, przedstawiona rozprawa dotyczy ważnego i aktualnego problemu badawczego związanego z efektywnością obliczeniową modeli głębokiego uczenia. Na szczególną uwagę zasługuje spójność tematyczna przedstawionych prac oraz fakt, że zostały one opublikowane w uznanych międzynarodowych konferencjach i czasopismach z zakresu uczenia maszynowego.

W mojej ocenie rozprawa spełnia wszystkie wymagania ustawowe i zwyczajowe stawiane pracom doktorskim i rekomenduję przyznanie magistrowi inżynierowi Bartoszowi Wójcikowi stopnia doktora na podstawie przedstawionej rozprawy.

Marek Cygan

Marek Cygan Marek Cygan