dr hab. Marek Śmieja, prof. UJ
Instytut Informatyki i Matematyki Komputerowej
Wydział Matematyki i Informatyki
Uniwersytet Jagielloński
marek.smieja@uj.edu.pl

Kraków, 6.01.2026

Recenzja rozprawy doktorskiej „Improving Performance of Mixture of Experts Large Language Models" autorstwa mgr Jana Ludziejewskiego

Tematyka rozprawy

Rozprawa doktorska poświęcona jest zagadnieniom zwiększania efektywności obliczeniowej i skalowalności dużych modeli językowych (Large Language Models, LLM). Kluczowym narzędziem jest architektura typu Mixture of Experts (MoE), w której – w zależności od przekazanego wejścia – aktywowana jest jedynie część modelu. Tematyka ta jest niezwykle aktualna i istotna zarówno z punktu widzenia badań podstawowych, jak i praktycznych zastosowań modeli językowych, których trening i wdrażanie wiążą się z bardzo wysokimi kosztami obliczeniowymi i pamięciowymi.

Autor koncentruje się na problemach skalowania modeli, formułowaniu praw skalowania (scaling laws), optymalnym doborze hiperparametrów oraz projektowaniu nowych architektur, które pozwalają uzyskać lepszy kompromis pomiędzy jakością modeli a wymaganiami sprzętowymi. Szczególną uwagę poświęcono architekturom rzadkim, w których tylko część parametrów jest aktywna dla danego tokenu, co wpisuje się w dominujący obecnie kierunek badań nad wydajnymi LLM.

Zawartość rozprawy

Rozprawa została złożona w tradycyjnej formie jako osobny dokument (nie jako cykl publikacji). Składa się ze wstępu, konkluzji oraz pięciu rozdziałów zawierających wyniki badań. Mimo że rozprawa nie jest formalnie cyklem publikacji, każdy z pięciu rozdziałów został napisany na podstawie osobnej pracy naukowej, której autor rozprawy jest jednym ze współautorów. Trzy z tych prac są artykułami opublikowanymi w czołowych konferencjach z dziedziny uczenia maszynowego (ICML, NeurIPS), a dwie pozostałe mają formę preprintów.

Poszczególne rozdziały zostały oparte na wieloautorskich pracach naukowych, co oznacza, że autor nie miał pełnego wkładu we wszystkie wyniki prezentowane w rozprawie. Trudno jednoznacznie ocenić wkład autora w powstanie rozprawy na podstawie przedstawionej

dokumentacji. Rozdział nr 1 zawiera określenie procentowego udziału w powstaniu kolejnych prac, który waha się w granicach od 10% do 40%. Brakuje jednak merytorycznego opisu wskazującego, co dokładnie stanowi wynik pracy mgr. Jana Ludziejewskiego. Taki opis umożliwiłby rzetelniejszą ocenę rozprawy z punktu widzenia doktoranta.

Poniżej przedstawiam zawartość każdego z pięciu rozdziałów [R2–R6] wraz z krótkim podsumowaniem.

[R2] Scaling Laws for Fine-Grained Mixture of Experts

Rozdział oparty jest na publikacji z konferencji ICML 2024, która ma 12 współautorów; autor rozprawy jest pierwszym autorem i deklaruje swój wkład na poziomie 23%.

Autor skupia się na modyfikacji architektury Mixture of Experts (MoE), aby efektywnie wykorzystać ją w scenariuszu obliczeń warunkowych. Idea polega na redukcji wymiarowości warstwy ukrytej w MoE w stosunku do typowej warstwy feed-forward. Stosunek wymiarowości warstwy feed-forward do wymiarowości analogicznej warstwy w MoE został nazwany przez autora ziarnistością (granularity, G). Zwiększając ziarnistość, token może być kierowany do więcej niż jednego eksperta przy zachowaniu liczby FLOPs porównywalnej ze standardowym modelem. Pozwala to zwiększyć elastyczność modelu, co w konsekwencji przekłada się na jego jakość. Następnie autor wyprowadza prawa skalowania, uwzględniając wprowadzony parametr ziarnistości oraz długość treningu. Stanowi to rozszerzenie analiz przeprowadzonych w wcześniejszych pracach. Wykorzystując zaproponowane prawa skalowania, autor pokazuje, że modele MoE, użyte w sposób optymalny, przewyższają klasyczne transformatory przy zadanym budżecie obliczeniowym, co kwestionuje wcześniejsze wyniki literaturowe.

Na wysoką ocenę zasługuje zarówno strona teoretyczna rozdziału, jak i jego bogate zaplecze empiryczne. Zaproponowane prawa skalowania zostały starannie dopasowane do danych i poddane procedurom walidacyjnym. Wyniki prowadzą do mocnej i dobrze uargumentowanej tezy, że odpowiednio skonfigurowane modele MoE są zawsze bardziej efektywne obliczeniowo niż modele gęste, niezależnie od skali. Analiza koncentruje się jednak głównie na miarach strat treningowych, pozostawiając otwartą kwestię wpływu ziarnistości na stabilność uczenia (np. zapadanie się ekspertów) oraz jakościową specjalizację ekspertów (w tym interpretowalność).

[R3] Joint MoE Scaling Laws: Mixture of Experts Can Be Memory Efficient

Rozdział oparty jest na publikacji z konferencji ICML 2025, która ma 9 współautorów; autor rozprawy jest pierwszym autorem i deklaruje swój wkład na poziomie 40%.

Rozdział trzeci w naturalny sposób rozwija wcześniejsze rozważania, rozszerzając analizę skalowania o jednoczesne uwzględnienie ograniczeń obliczeniowych i pamięciowych. Autor proponuje metodologię obejmującą zarówno modele gęste, jak i architektury MoE, co pozwala na bezpośrednie porównanie tych dwóch paradygmatów. Szczególnie istotnym wynikiem jest wykazanie, że wbrew powszechnym przekonaniom modele MoE mogą być korzystne również z punktu widzenia zużycia pamięci, a nie tylko kosztów obliczeniowych.

Podobnie jak w przypadku poprzedniego rozdziału, na uwagę zasługuje odwaga autora, który kwestionuje wcześniejsze przekonania dotyczące kosztów pamięciowych modeli MoE. Rozdział charakteryzuje się dużą starannością metodologiczną. Autor precyzyjnie definiuje różne pojęcia

optymalności (obliczeniowej, pamięciowej, inferencyjnej) i konsekwentnie analizuje je w ramach jednego modelu matematycznego. Praca ma wyraźny walor aplikacyjny, dostarczając projektantom dużych modeli konkretnych wskazówek dotyczących kompromisów pomiędzy jakością a kosztami sprzętowymi. Wprowadzony model pamięciowy nie uwzględnia jednak w pełni kosztów aktywacji oraz buforów pośrednich w realistycznych implementacjach, co stanowi pewne ograniczenie wyników.

[R4] Different Rates for Different Weights: Decoupled Relative Learning Rate Schedules

Rozdział oparty jest na preprincie, który ma 11 współautorów; autor rozprawy jest pierwszym autorem i deklaruje swój wkład na poziomie 40%.

Rozdział czwarty poświęcony jest zagadnieniom optymalizacji procesu uczenia i stanowi wyraźne odejście od czysto architektonicznych rozważań wcześniejszych części rozprawy. Autor proponuje metodę, w której różnym komponentom modelu przypisuje się odmienne tempa uczenia, przy jednoczesnym zachowaniu ich wzajemnych proporcji podczas skalowania modelu. Aby zachować wysoką efektywność obliczeniową, dobór hiperparametrów przeprowadzany jest na zmniejszonych modelach, gdzie koszt uczenia jest znacznie niższy. Koncepcja ta jest prosta i elegancka, a jednocześnie dobrze umotywowana empirycznie.

Największą zaletą rozdziału jest wykazanie wysokiej skalowalności zaproponowanej metody. Autor pokazuje, że hiperparametry dobrane dla małych modeli mogą być skutecznie przenoszone na modele wielokrotnie większe, co ma ogromne znaczenie praktyczne. Uzyskane przyspieszenia treningu, sięgające kilkudziesięciu procent, osiągnięto bez pogorszenia stabilności czy jakości modeli. Wyniki te świadczą o bardzo dobrym wyczuciu praktycznych problemów związanych z trenowaniem LLM. Rozdział ma charakter empiryczny, natomiast teoretyczne uzasadnienie obserwowanych zależności pozostaje ograniczone.

[R5] Mixture of Tokens: Efficient LLMs through Cross-Example Aggregation

Rozdział oparty jest na publikacji z konferencji NeurIPS 2024, która ma 10 współautorów; autor rozprawy jest piątym autorem i deklaruje swój wkład na poziomie 10%.

Model przedstawiony w rozdziale R5 modyfikuje mechanizm sterowania w klasycznym MoE, prowadząc do jego wersji ciągłej. Idea polega na agregowaniu informacji zawartej w grupach tokenów, a następnie przekierowywaniu jej do wszystkich ekspertów, po czym wyniki są dekodowane do pojedynczych tokenów. Ponieważ eksperci nie przetwarzają już pojedynczych tokenów, każdy z nich może być wykorzystany do przetwarzania tokenu wynikowego bez zwiększania kosztu obliczeniowego. Dodatkowo autor usprawnia metodę poprzez zastosowanie większej liczby wysokoziarnistych ekspertów oraz odpowiednie grupowanie tokenów podlegających agregacji.

Wyniki pokazują, że opracowany model prowadzi do znaczącego przyspieszenia względem klasycznego transformera przy zachowaniu jakości predykcji. Pewnym ograniczeniem jest wzrost zużycia pamięci wynikający z zastosowania warstw MoE. Ponadto mieszanie tokenów z różnych sekwencji uniemożliwia przetwarzanie pojedynczych przykładów.

[R6] MoE-Mamba: Efficient Selective State Space Models with Mixture of Experts

Rozdział oparty jest na preprincie, który ma 10 współautorów; autor rozprawy jest czwartym autorem i deklaruje swój wkład na poziomie 15%.

Rozdział szósty rozszerza zakres rozprawy poza architektury transformerowe, łącząc ideę Mixture of Experts z modelami typu State Space Models, w szczególności z architekturą Mamba. Autor sprawdza eksperymentalnie, czy MoE może być z powodzeniem zastosowane w tego typu modelach. Wyniki pokazują, że MoE również w tym przypadku prowadzi do poprawy wydajności, co sugeruje, że jest to technika o charakterze uniwersalnym, możliwa do zastosowania w różnych architekturach.

Na uwagę zasługuje zachowanie korzystnych właściwości inferencyjnych modeli SSM przy jednoczesnym zwiększeniu efektywności treningu. Rozdział ten pokazuje dużą elastyczność myślenia autora oraz zdolność do przenoszenia idei pomiędzy różnymi klasami modeli. Jest to istotny sygnał, że zaproponowane w rozprawie koncepcje mają potencjał wykraczający poza jeden dominujący obecnie paradygmat.

Ograniczeniem rozdziału jest koncentracja na jednej konkretnej architekturze SSM oraz brak szerszych porównań z innymi nowoczesnymi alternatywami dla transformerów. Część wniosków ma charakter wstępny i wymaga dalszych badań.

Dyskusja

Ogólne wrażenie po lekturze rozprawy jest pozytywne. Autor podjął temat o fundamentalnym znaczeniu, znajdujący się w centrum zainteresowania wiodących ośrodków badawczych na świecie. Jednocześnie potrafił przebić się ze swoimi wynikami na najlepszych konferencjach z zakresu uczenia maszynowego. Trzy publikacje na konferencjach klasy A* w ciągu czterech lat są osiągnięciem wybitnym jak na doktoranta, a także bardzo znaczącym z punktu widzenia doświadczonego naukowca. Dane dotyczące cytowalności z bazy Google Scholar (ponad 200 cytowań) również pozytywnie świadczą o dorobku doktoranta oraz randze jego wyników. Należy podkreślić, że tematyka rozprawy jest spójna, co nie jest łatwe do osiągnięcia w przypadku doktoratów opartych na artykułach lub preprintach. W działalności doktoranta widoczna jest wyraźna myśl przewodnia, która konsekwentnie prowadzi do uzyskanych rezultatów. Wszystkie prace, na podstawie których powstała rozprawa, są pracami wieloautorskimi, a liczba współautorów oscyluje wokół dziesięciu. Biorąc jednak pod uwagę specyfikę badań nad dużymi modelami, należy zaakceptować taki charakter współpracy. Pewne wątpliwości budzi fakt, że trzy z pięciu prac zostały wcześniej uwzględnione w rozprawie doktorskiej dr. Sebastiana Jaszczura, który jest jednocześnie współautorem wszystkich pięciu prac. Brakuje w rozprawie precyzyjnego określenia indywidualnego wkładu doktoranta, co utrudnia jednoznaczną ocenę jego osiągnięć.

Rozprawa charakteryzuje się bardzo wysokim poziomem merytorycznym i metodologicznym. Na szczególne uznanie zasługuje umiejętne połączenie solidnych podstaw teoretycznych (nowe prawa skalowania) z szeroko zakrojonymi eksperymentami empirycznymi, przeprowadzonymi na dużą skalę. Autor nie ogranicza się do jednego aspektu problemu, lecz konsekwentnie analizuje zarówno architekturę modeli, jak i proces ich uczenia oraz ograniczenia sprzętowe. Istotną zaletą pracy jest krytyczne odniesienie się do wcześniejszych wyników literaturowych i

wykazanie, że niektóre pesymistyczne wnioski dotyczące skalowania MoE wynikają z przyjęcia zbyt restrykcyjnych założeń. Zaproponowane uogólnienia i nowe parametry (takie jak ziarnistość) znacząco pogłębiają zrozumienie zachowania modeli rzadkich. Praca jest napisana jasno i precyzyjnie, choć ze względu na zaawansowany charakter zagadnień wymaga od czytelnika bardzo dobrej znajomości współczesnych metod uczenia głębokiego. Pewnym ograniczeniem jest fakt, że część rozdziałów ma charakter rozwiniętych artykułów konferencyjnych, co miejscami prowadzi do nierównomiernego poziomu szczegółowości.

Na zakończenie dyskusji pozwolę sobie sformułować kilka pytań:

1. **Granice obowiązywania praw skalowania.** Czy autor dostrzega empiryczne lub teoretyczne przesłanki wskazujące, że zaproponowane prawa skalowania dla MoE mogą przestać obowiązywać przy jeszcze większych skalach modeli lub w innych reżimach danych (np. bardzo małych lub silnie zaszumionych zbiorach)?

2. **Specjalizacja ekspertów.** Czy autor obserwował jakościowe różnice w specjalizacji ekspertów pomiędzy klasycznymi MoE, MoT oraz MoE-Mamba i czy możliwe jest wykorzystanie tej specjalizacji do poprawy interpretowalności modeli?

3. **Znaczenie wyników dla praktycznych systemów LLM.** Który z zaproponowanych elementów (nowe prawa skalowania, RLRS, MoT, MoE-Mamba) autor uważa za najbardziej perspektywiczny z punktu widzenia wdrożeń przemysłowych w najbliższych latach i dlaczego?

4. **Ogólność rozwiązań.** W rozdziale R6 pokazano, że MoE może być wykorzystane w modelu Mamba. Sugeruje to, że wyniki uzyskane w rozprawie mogą znaleźć zastosowanie również w innych modelach. Jak autor ocenia możliwość transferu tych wyników do jeszcze innych architektur?

Konkluzja

Przedłożona rozprawa mgr. Jana Ludziejewskiego zatytułowana *"Improving Performance of Mixture of Experts Large Language Models"* stanowi znaczący i oryginalny wkład w rozwój badań nad dużymi modelami językowymi oraz architekturami Mixture of Experts. Autor zaproponował nowe koncepcje teoretyczne, metody optymalizacji oraz innowacyjne rozwiązania architektoniczne, które mają realne znaczenie dla dalszego rozwoju wydajnych modeli językowych.

Zakres i jakość uzyskanych wyników, ich solidne potwierdzenie eksperymentalne oraz publikacje w prestiżowych miejscach jednoznacznie świadczą o dojrzałości naukowej autora. Rozprawa spełnia wszystkie wymagania ustawowe oraz zwyczajowe stawiane pracom doktorskim w dyscyplinie informatyka i może być z pełnym przekonaniem rekomendowana do dopuszczenia do dalszych etapów przewodu doktorskiego.

Podpisany elektronicznie przez
Marek Andrzej Śmieja
08.01.2026
15:43:02 +01'00'