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

Kraków, 17.01.2025

Recenzja rozprawy doktorskiej
„Efficient Large Language Models with Conditional Computation"
autorstwa mgr Sebastiana Jaszczura

Tematyka rozprawy

Rozprawa doktorska koncentruje się wokół dużych modeli językowych (LLMy), które w większości realizowane są w oparciu o architekturę transformera. Z uwagi na fakt, że LLMy zawierają ogromną liczbę parametrów oraz są trenowane z użyciem gigantycznych zbiorów danych, to kluczową kwestią jest efektywna, pod względem obliczeniowym, realizacja tych modeli. W celu usprawnienia działania LLMów autor rozważa w przeważającej części podejścia oparte o tak zwane obliczenia warunkowe, gdzie sieć decyduje jaka jej część będzie przetwarzać dane wejście. Aktywując jedynie wybrane fragmenty sieci, możemy znacząco zmniejszyć koszt obliczeniowy związany z przetwarzaniem danego przykładu. Aby to podejście mogło być stosowane w praktyce, to model nie może obniżać swoich zdolności do rozwiązania zadania, do którego został zaprojektowany. Wszystkie zaprezentowane metody, poza jedną, są aplikowane do architektury transformera. Jako narzędzie umożliwiające zastosowanie obliczeń warunkowych, autor wybrał architekturę *mixture of experts,* która jest rozważana w większości rozdziałów.

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 rozprawy. Mimo, że rozprawa nie jest formalnie cyklem publikacji, to każdy z pięciu rozdziałów jest napisany na podstawie osobnej pracy naukowej, której autor rozprawy jest jednym z współautorów. Trzy z tych prac są artykułami opublikowanymi w czołowych konferencjach z dziedziny uczenia maszynowego (NeurIPS, ICML), a dwie

pozostałe to preprinty. Ponieważ rozdziały oparte zostały o wieloautorskie prace naukowe, to autor rozprawy nie miał pełnego wkładu w uzyskanie wyników. Autor deklaruje swój wkład w rozdziale pierwszym.

Poniżej przedstawię zawartość każdego z pięciu rozdziałów [R2-R6] wraz z krótkim podsumowaniem:

[R2] Sparse is Enough in Scaling Transformers

Rozdział oparty jest na publikacji z konferencji NeurIPS 2021, która ma 7 współautorów, a autor rozprawy jest autorem wiodącym i deklaruje swój wkład na 40%.

Autor stawia sobie za cel przyspieszenie inferencji w modelach transformerowych. Do tego celu proponuje rzadkie wersje wszystkich warstw transformera, włączając w to warstwę *feed-forward* oraz warstwy atencyjne. W przypadku sieci *feed-forward,* autor zauważa, że jej jedyna ukryta warstwa tworzy reprezentację zawierającą wiele zer, z uwagi na zastosowanie aktywacji ReLU. Idąc o krok dalej autor narzuca ustaloną rzadką strukturę tej reprezentacji, która będzie sterowana przez tak zwany kontroler. Dla każdego wejścia, kontroler decyduje o kształcie tej reprezentacji, co realizuje paradygmat obliczeń warunkowych. Zakładając ustaloną strukturę reprezentacji, część obliczeń może być wykonywana efektywnie, bo nie będzie już mnożeniem macierzy o pełnym wymiarze. Dla realizacji atencji autor rozbija jej operacje na oddzielne komponenty, które można efektywnie zrealizować zakładając jej rzadką reprezentację. W każdym przypadku autor podaje spadek liczby operacji w stosunku do standardowego transformera. Dodatkowo autor zauważa, że koszt przetwarzania długich ciągów przez bloki atencyjne może zdominować pozostałe operacje. Aby temu zapobiec wykorzystywany jest rzadka wersja atencji (Locality-Sensitive Hashing).

Przeprowadzone eksperymenty pokazują, że zaproponowane rzadkie wersje modeli dorównują tradycyjnym transformerom pod względem jakości, a przy tym są znacząco szybsze w czasie inferencji. Szybkość jest tutaj mierzona ilością operacji zmiennoprzecinkowych (FLOPs), a nie faktycznym czasem obliczeń. Należy zauważyć (co autor robi), że nie wszystkie operacje są dobrze wspierane przez obecny sprzęt. W konsekwencji mniejsza liczba operacji liczona dla jednego przykładu, może w rzeczywistości nie dać zysku czasowego w przypadku przetwarzania całej porcji danych. Nie mniej jednak rozdział jest bardzo ciekawy i stoi na wysokim poziomie. Mimo, że autor nie używa zaawansowanych narzędzi, to widoczne jest bardzo dobre zrozumienie szczegółów w złożonym modelu, jakim jak transformer, co pozwala na znaczącą optymalizację jego działania.

[R3] Scaling Laws for Fine-Grained Mixture of Experts

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

Autor skupia się na modyfikacji modelu mixture of experts (MoE), aby efektywnie ją wykorzystać 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 nazwana jest przez autora *grannularnością* (G). Zwiększając granularność, token może być kierowany do więcej niż jednego eksperta jeśli chcemy zachować ilość FLOPsów analogiczną do standardowego modelu. Pozwala to zwiększyć elastyczność modelu, co w konsekwencji przełoży się na wynik modelu.

Następnie autor zajmuje się wyprowadzeniem praw skalowania biorąc pod uwagę wprowadzony parametr granularności, a także długość treningu. Rozszerza to analizy przeprowadzone przez autorów wcześniejszych prac. Używając przedstawionych praw skalowalności autor pokazuje, że modele z optymalnie użytym MoE zawsze przewyższają typowe transformery przy założonym budżecie obliczeniowym, co stoi w sprzeczności z wynikami uzyskanymi wcześniej w literaturze.

Autor zauważa, że w odróżnieniu od metody z rozdziału R2, wprowadzona technika może być zoptymalizowana na GPU. Podczas gdy metoda z rozdziału R2 dotyczyła maksymalnej granularności, a standardowy model MoE jest przypadkiem minimalnej granularności, to technika z tego rozdziału pozwala osiągnąć granularności pośrednie. Wyniki uzyskane w tym rozdziale były nieoczywiste, tym bardziej że częściowo kwestionują powszechnie stosowane podejścia.

[R4] Mixture of Tokens: Efficient LLMs through Cross-Example Aggregation

Rozdział oparty jest na publikacji z konferencji NeurIPS 2024, która ma 10 współautorów, a autor rozprawy jest ostatnim autorem i deklaruje swój wkład na 25%.

Model przedstawiony w rozdziale R3 może być traktowany jako model nieciągły, ponieważ informacja jest kierowana do wybranych ekspertów, co powoduje niestabilność uczenia. Model przedstawiony w rozdziale R4 zmienia sterowanie w klasycznym MoE prowadząc do wersji ciągłej. Idea polega na zagregowaniu informacji zawartej w grupie tokenów, a następnie przekierowaniu jej do wszystkich ekspertów, co ostatecznie jest dekodowane do pojedynczych tokenów. Ponieważ eksperci nie przetwarzają już pojedynczych tokenów, to każdy z nich może być użyty do przetwarzania tokenu wynikowego bez zwiększenia kosztu obliczeniowego. Dodatkowo, autor usprawnia metodę używając więcej wysoko-granularnych ekspertów, a także odpowiednio grupując tokeny podlegające agregacji.

Wyniki przedstawione w tym rozdziale pokazują, że opracowany model powoduje znaczące przyspieszenie w stosunku do klasycznego transformera zachowując jakość predykcji. Pewną wadą modelu jest wzrost użycia pamięci związany z użyciem warstw MoE. Ponadto, mieszanie tokenów z różnych sekwencji uniemożliwia przetwarzania pojedynczych przykładów.

[R5] Structured Packing in LLM Training Improves Long Context Utilization

Rozdział oparty jest na preprincie, który ma 7 współautorów, a autor rozprawy jest trzecim autorem i deklaruje swój wkład na 15%.

Ten rozdział różni się tematycznie od poprzednich. Autor rozważa problem przetwarzania długiego kontekstu w modelach transformerowych. O ile modele językowe radzą sobie zwykle z przetwarzanie długich dokumentów (na przykład książek), to jednak ich efektywność spada w tym przypadku. Zamiast ingerować w architekturę modelu, autor skupia się na odpowiednim przygotowaniu danych. Idea podejścia polega na organizacji dokumentów w postaci drzewa na podstawie ich podobieństwa. Dzięki temu do modelu wprowadzana jest dodatkowa wiedza dotycząca podobieństwa dokumentów. Autor rozważa kilka technik wydobywania dokumentów podobnych. Eksperymenty pokazują, że taka technika poprawia wyniki na docelowych zadaniach, a także wpływa na szybkość treningu.

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

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

Motywacją do napisania tego rozdziału jest spopularyzowanie nowego modelu Mamba, który stanowi obecnie alternatywę do transformera. Autor sprawdza eksperymentalnie czy MoE może być z powodzeniem zastosowane w przypadku tej architektury. Wyniki pokazują, że MoE również wpływa na poprawę wydajności tego modelu, co prowadzi do hipotezy, że MoE jest uniwersalną techniką mogącą towarzyszyć różnym architekturom.

Jest to najmniej nowatorski rozdział w całej rozprawie. Nie mniej jednak pokazuje, że autor doskonale zna i potrafi się posługiwać najnowszymi modelami językowymi, a także analizować ich działanie.

Dyskusja

Ogólne wrażenie po przeczytaniu rozprawy jest bardzo pozytywne. Autor podjął „temat na czasie", który jest w centrum zainteresowania najważniejszych ośrodków na świecie. Potrafił jednocześnie przebić się ze swoimi wynikami na najlepszych konferencjach z uczenia maszynowego. Trzy publikacje na konferencjach A* w ciągu 4 lat są wynikiem wybitnym jak na doktoranta, a również znaczącym z punktu widzenia doświadczonego naukowca. Dane dotyczące cytowalności z bazy google scholar (ponad 200 cytowań) również pozytywnie świadczą o dorobku doktoranta. Należy podkreślić, że tematyka rozprawy jest dość spójna, co nie jest łatwe do uzyskania w przypadku doktoratów opartych na artykułach lub preprintach. W działaniach doktoranta widoczna jest głębsza myśl przewodnia, która doprowadziła go do końcowych wyników.

Wszystkie prace, na podstawie których powstała rozprawa, są pracami wieloautorskimi, a liczba współpracowników oscyluje wokół liczby 10. Mając jednak świadomość jak wygląda praca z dużymi modelami, należy zaakceptować mnogość współautorów. Próbując zmniejszać liczbę współpracowników, badania z pewnością znacząco by się wydłużyły, co mogłoby doprowadzić do dezaktualizacji uzyskiwanych wyników. Na wstępie rozprawy autor podkreśla, że poza wiodącą rolą w uzyskaniu większości wyników, brał również udział w kierowaniu pracami, a także tworzył i kierował zespołem mniej doświadczonych współpracowników. Nie mam wątpliwości, że autor potwierdził swoją samodzielność w prowadzeniu badań, jak również pokazał, że potrafi krytycznie podejść do badanych problemów.

Praca jest dobrze zredagowana. Dużym ułatwieniem jest kończenie każdego rozdziału konkluzją, która pokazuje pewne wady opracowanego podejścia, co następnie jest adresowane w kolejnym rozdziale. Pewną trudnością może być jednak brak wytłumaczenia podstaw dotyczących modeli językowych i transformerów. Do pełnego zrozumienia wyników konieczna jest bardzo dobra znajomość specyficznej dziedziny sztucznej inteligencji, czego jednak autor nie stara się przedstawić. Dla przykładu, jako podstawową miarę autor używa perplexity, ale w żadnym miejscu nie tłumaczy co oznacza (z tekstu czytelnik mógłby się nawet nie zorientować czy zależy nam na maksymalizacji czy minimalizacji tej miary). O ile można by częściowo założyć, że ta miara jest znana dla starszego typu modeli językowych, o tyle w przypadku LLMów nie jest to oczywiste. Podobna sytuacja ma miejsce w innych fragmentach pracy np. co oznaczają miary w tabeli 2.7? Te minusy nie przewyższają jednak ogólnego dobrego wrażenia przeczytanej pracy.

Na koniec dyskusji pozwolę sobie zadać kilka pytań:

1. MoE zostało przez autora użyte w celu optymalizacji obliczeń. Skoro każdy ekspert specjalizuje się w przetwarzaniu dedykowanego wejścia, to czy możemy wyciągnąć z tego modelu jakąś informację dotyczącą interpretowalności? Ewentualnie, czy zastosowanie MoE daje nam jeszcze dodatkowe korzyści poza obliczeniami warunkowymi?
2. Często w praktyce pracujemy z pre-trenowanymi modelami. Czy autor widzi sens destylacji typowego modelu transformera do postaci z MoE i jak taka destylacja mogłaby wyglądać?
3. Autor wspomina, że obliczenia warunkowe często nie mogą być zoptymalizowane na kartach graficznych. Jak w dłuższej perspektywie wygląda potencjalne rozwiązanie tego problemu? Czy jest szansa, aby przezwyciężyć ten problem i czy obliczenia warunkowe mają szanse stać się popularne w rozwiązaniach praktycznych?
4. W rozdziale R6 pokazano, że MoE może być wykorzystane w modelu Mamba. Sugeruje to, ze wyniki uzyskane w rozprawie mogą być zastosowane również w

innych modelach. Czy wyniki z rozdziałów 2-4 można by przenieść na inne typowe architektury sieci gęstych, czy konwolucyjnych, i czy mogłoby to doprowadzić do ich usprawnienia?

Konkluzja

Przedłożona rozprawa mgr Sebastiana Jaszczura zatytułowana „Efficient Large Language Models with Conditional Computation" (pol. Efektywne duże modele językowe przy użyciu obliczeń warunkowych) stanowi istotne osiągnięcie naukowe w dyscyplinie informatyka, a autor w jej przygotowaniu wykazał się samodzielnością, pomysłowością i bardzo dobrym opanowanym warsztatem najnowszych narzędzi sztucznej inteligencji. Jestem przekonany, że spełnia ona wymagania ustawowe i zwyczajowe stawiane pracom doktorskim i może stanowić podstawę do nadania stopnia doktora. Co więcej, oceniam, że przedstawiona rozprawa wykracza pozytywnie poza zwykłe wymagania stawiane rozprawom doktorskim w informatyce. Dlatego też wnioskuję o wyróżnienie rozprawy.

Podpisany elektronicznie przez
Marek Andrzej Śmieja
17.01.2025
15:15:36 +01'00'