Wydajność i Adaptacyjność Poprzez Kompresję Dużych Modeli Językowych

Streszczenie popularnonaukowe

W ostatnich latach szybki rozwój dużych modeli językowych (LLM) przyczynił się do istotnych postępów w zakresie rozumienia i generowania języka naturalnego. Rozwój ten wiąże się jednak z ogromnymi kosztami obliczeniowymi i pamięciowymi zarówno podczas treningu, jak i wdrażania modeli. Pomimo imponujących możliwości, najnowocześniejsze modele LLM pozostają trudne do odtworzenia, szeroko niedostępne i wysoce obciążające środowisko naturalne ze względu na swój rozmiar. Rośnie pilna potrzeba opracowania skutecznych strategii kompresji, które nie tylko ułatwią wdrażanie, ale także ułatwią badania poza dużymi ośrodkami przemysłowymi. Wyzwanie to dobrze ilustruje nowo wybudowane centrum obliczeniowe xAI przeznaczone do trenowania modeli sztucznej inteligencji (AI) nowej generacji. Szacuje się, że sama ta infrastruktura zużywa codziennie setki megawatogodzin energii elektrycznej - ilość wystarczającą do zasilenia dziesiątek tysięcy gospodarstw domowych. Skala tego zużycia energii podkreśla, jak ważne jest, by modele AI były nie tylko coraz inteligentniejsze, ale również bardziej wydajne.

Aby stawić czoła temu wyzwaniu, nasze badania skoncentrują się na lepszym zrozumieniu mechanizmów kompresji oraz ich wpływu na prawa skalowania efektywnego trening dużych modeli językowych. Celem jest opracowanie lepszych metod kompresji oraz przeprowadzenie badań, które pozwolą głębiej zrozumieć kompromisy za nimi idące pomiędzy jakością a wydajnością modeli. Proponowany projekt przyczyni się do stworzenia bardziej dostępnej i zrównoważonej przyszłości dla dużych modeli językowych. W jego wyniku powstaną metody i wnioski przydatne zarówno dla środowisk akademickich, jak i przemysłowych zajmujących się optymalizacją modeli, ich interpretowalnością oraz rozwojem zrównoważonej sztucznej inteligencji.

Projected Compression

Projected Compression to nowa metoda, która uzupełnia standardowe podejścia do przycinania wag modelu, wykorzystując wszystkie parametry redukowanego modelu podczas retreningu. W przeciwieństwie do konwencjonalnych, Projected Compression zachowuje dostęp do wszystkich parametrów modelu bazowego poprzez uczące się moduły projekcyjne. Umożliwia to odzyskiwanie poprzez metody gradientowej optymalizacji informacji z oryginalnego modelu przy jednoczesnym zachowaniu efektywności i zgodności standardowych rozwiązań. Integracja wszystkich parametrów modelu bazowego podczas retreningu ma potencjał znacząco zwiększyć efektywność kompresji, wykorzystując pełną reprezentację bazowego modelu.

Kompresja Przestrzeni Logitów LLM poprzez Redukcję Wymiarowości w Procesie Destylacji

Podczas destylacji większego modelu do mniejszego wyjściowe prawdopodobieństwa tokenów (logity) zawierają bogate informacje semantyczne, ich pełna wymiarowość (zazwyczaj przekraczająca 30000 tokenów) wiąże się ze znacznymi kosztami pamięciowymi i komunikacyjnymi. Aby je ograniczyć, proponujemy kompresję logitów pełnego słownika z wykorzystaniem technik takich jak PCA. Destylacja w niskowymiarowej przestrzeni logitów, z dynamicznym ich rozszerzaniem, pozwala zmniejszyć zużycie zasobów sprzętowych przy zachowaniu wysokiej jakości modelu. Dodatkowo, analizując wpływ poszczególnych wymiarów logitów na proces uczenia modelu ucznia, zyskujemy wgląd w to, które tokeny niosą najbardziej transferowalną wiedzę, co usprawni zarówno interpretowalność, jak i sam proces destylacji.

Prawa skalowania dla hybrydowych metod kompresji dużych modeli językowych

Aby wspierać praktyczne decyzje dotyczące kompresji modeli w różnych zastosowaniach i ograniczeniach sprzętowych, planujemy zbadać prawa skalowalności metod hybrydowej kompresji, które łączą techniki takie jak przycinanie, kwantyzacja czy destylacja. Choć różne metody kompresji oferują pozytywne kompromisy, ich wzajemne oddziaływania są słabo poznane. Nasze badania będą obejmować zarówno eksperymenty empiryczne, jak i modelowanie teoretyczne ekstrapolacji, w celu ustalenia, jak zmienia się wydajność treningu wraz z rozmiarem modelu, jego jakością, poziomem redukcji oraz zastosowanymi metodami kompresji. Celem jest wyprowadzenie uogólnionych praw skalowania, które będą wspierać wybór optymalnej strategii kompresji w procesie rozwoju dużych modeli językowych.