Dostrajanie LLM: Dostrajanie z Efektywnym Dostrajaniem Instrukcyjnym i Parametrowym (PEFT)

Streszczenie

W funkcjach generujących sztuczną inteligencję (AI) model dużego języka odgrywa ważną rolę w różnorodnej komunikacji długich danych. Mimo że system posiada wstępnie wytrenowany model języka, ten model dużego języka (LLM) będzie wspomagał model zgodnie z wytrenowanymi danymi w różnych analizach. System ten jest bardzo przydatny w wielu dziedzinach, takich jak przetwarzanie języka naturalnego, odpowiadanie na pytania i GPT, w celu uzyskania lepszej wydajności. Gromadzenie dużych ilości danych w wstępnie wytrenowanym modelu zapewni lepsze rozwiązanie, ale nie jest w stanie dostroić wszystkich danych, co jest odpowiednie dla aplikacji wielozadaniowych. Ten model LLM służy do efektywniejszego trenowania danych zgodnie z różnymi kategoriami, takimi jak predykcja zbiorów danych medycznych, tłumaczenie językowe i wsparcie użytkownika. Aby zwiększyć poziom dokładności, stosowana jest metoda wytrenowanego dostrajania, która jest bardziej odpowiednia dla każdego konkretnego zadania. Wstępnie wytrenowany model służy do trenowania całego zbioru danych, który jest dostępny w danym zadaniu i analizowany za pomocą precyzyjnie dostrojonych przejść danych. Chociaż ten wstępnie wytrenowany model generuje dużą analizę danych w zakresie przetwarzania języka, szkolenia języka i predykcji za pomocą LLM, zapewnia on dobrą wydajność dla dowolnego zbioru danych. Model dostrajania służy do zrozumienia transmisji danych, predykcji zadań i analizy medycznej. Na przykład, trenowanie zbioru danych za pomocą dostrajania służy do efektywnego generowania przejść między dużymi zbiorami danych w różnych aplikacjach, takich jak LLM i GPT. Ten duży model językowy jest jedną z najlepszych sekwencji trenowanych zbiorów danych w zastosowaniach generatywnej sztucznej inteligencji w czasie rzeczywistym. Ten model LLM wie, jak zebrać zbiór danych zgodnie z wymaganiami użytkownika i jak wygenerować wynik za pomocą metody dostrajania, ale wymagał większej przestrzeni dyskowej w porównaniu z precyzyjnie dostrojonym modelem LLM. Na przykład, precyzyjne dostrajanie modelu LLM, takiego jak GPT, obejmuje pewien wytrenowany język zgodnie z predefiniowanym modelem; dzięki temu modelowi LLM, obecnie wytrenowany zbiór danych będzie analizował pytania i odpowiedzi w oparciu o obecną sytuację, ale występują pewne trudności ze zwiększeniem liczby zbiorów danych w przypadku zaawansowanych zastosowań. W tym przypadku do poprawnej analizy stosowana jest metoda precyzyjnego dostrajania z uwzględnieniem parametrów, uwzględniająca pojemność pamięci dla danego zadania.

Słowa kluczowe: Efektywność parametrów, precyzyjne dostrajanie, wstępnie wytrenowany model języka, duży model języka, wykorzystanie pamięci, GPT, AI, LSTM, CNN,

Wprowadzenie

Ten system ilustruje normalizację warstw w generatywnej sztucznej inteligencji (AI), która jest wykorzystywana w wielu zastosowaniach głębokiego uczenia się. W tym rozdziale przedstawiono różne modele treningowe sieci neuronowych. Ogólnie rzecz biorąc, sieci neuronowe mają pewne typy danych, które są już wytrenowane w oparciu o różne modele. Ten wytrenowany zbiór danych stanowi duże wyzwanie w podejściu normalizacyjnym, które służy do utrzymania stabilności danych, a także lepszej dokładności w ramach danej architektury. W dużym modelu językowym (LLM) różne architektury i zestawy danych odgrywają ważną rolę w ocenie zakresu konfiguracji zgodnie ze zbiorem danych językowych i wytrenowanym modelem. Proponowany system wprowadził architekturę neuronowego tłumaczenia maszynowego, która jest używana do modelowania języka zgodnie z wytrenowanym zestawem danych. Ta metoda jest używana do wysokiego przepływu danych w celu ulepszenia architektury. Wykorzystując tłumaczenie maszynowe i podejście informacyjne, system uzyska dokładne modele tłumaczeń. Sieć pamięci długoterminowej (LSTM) jest ważniejsza dla trenowania modelu w różnych aspektach. Maszyna może wykonać proces w oparciu o informacje

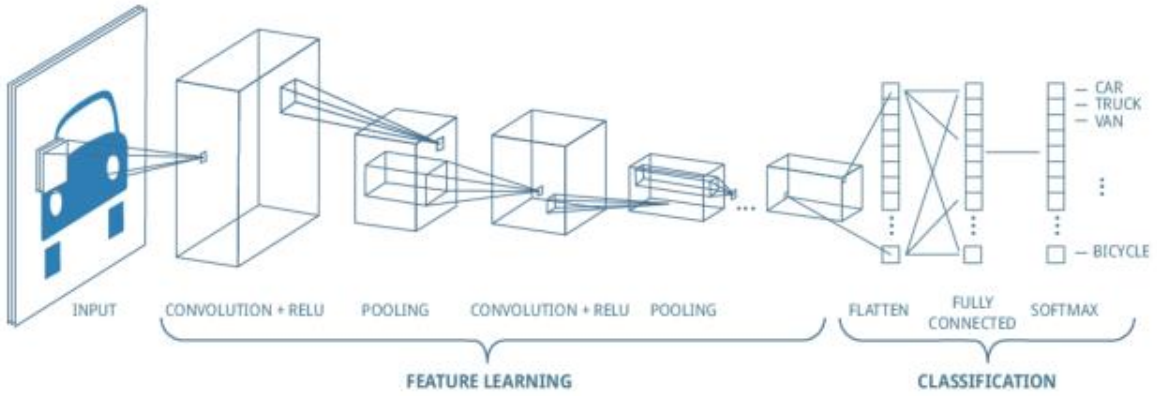
tekstowe, a system zidentyfikował pytania i odpowiedzi. W bardziej efektywny sposób LSTM jest używana do przechwytywania informacji o danych tekstowych przy użyciu modelu maszyny języka naturalnego w sieciach LSTM, a system przetwarza dane i buduje bardziej ustrukturyzowane czytanie i obsługę złożonych danych tekstowych. Prezentuje CNN, która jest konwencjonalną architekturą sieci neuronowej, która jest używana do minimalizacji złożoności obwodu dla dowolnej architektury. W istniejącej CNN ma większą liczbę wymagań systemowych w celu wdrożenia systemu lub architektury. W tradycyjnych sieciach CNN wydajność jest słaba z powodu opóźnień w architekturze sieci i metodzie splotu. W tym proponowanym systemie zaprojektowano przełomową sieć splotową, aby osiągnąć wysoką wydajność, mniejszą złożoność i większą dokładność. Ten typ warstwy splotu zapewnia odpowiednią długość kanału oraz zmniejszenie liczby parametrów i czasu obliczeniowego. Niniejszy rozdział przedstawia rekurencyjną sieć neuronową, która jest używana do przesyłania danych od końca do końca, co służy do redukcji błędów w zdaniach w oparciu o wytrenowany model. Sieć ta jest używana do podejścia z różnymi algorytmami w celu znalezienia przewidywanego zbioru danych. Większość sieci neuronowych działała w oparciu o zbiór danych i predefiniowany model. Proces uczenia maszynowego działa w oparciu o różne architektury i poziomy algorytmów. Ten typ sieci jest używany do trenowania i ulepszania ogólnych danych, co zapewnia większą dokładność dla LLM. Zgodnie z danymi wejściowymi, RNN wytrenował model, który zostanie zaimplementowany w dowolnym typie danych selekcji tekstu w RNN. Zaproponowane gramatyki rekurencyjnych sieci neuronowych odgrywają bardzo dużą rolę w podejściu do modelowania języka naturalnego. Ta odpowiedź gramatyczna w strukturze systemowej jest ważniejsza w przypadku analizy tekstu bezbłędnego. Jednak ta metoda wykorzystuje podejście oparte na języku naturalnym do optymalizacji parserów opartych na grafach. System ten posiada proces analizy składniowej w czasie rzeczywistym, który jest bardziej odpowiedni dla analizy RNNG. Nie przepada za istniejącymi sieciami neuronowymi (RNNN), a sieci neuronowe (RNNG) poprawią standard uczenia i generalizacji. System ten jest bardziej elastyczny i zapewnia lepszą generalizację dzięki wykorzystaniu sieci neuronowych (RRNG). System ten wykorzystuje architekturę sieci neuronowych typu sekwencja-sekwencja, która jest wykorzystywana do bardziej wymagających zadań. Ta metoda nie przepada za splotową siecią neuronową, która służy do przechwytywania całego zbioru danych zgodnie z wytrenowanym modelem. Metoda ta służy do poprawy skalowalności dzięki przejściu z sekwencji do sekwencji. Kolejnym ważnym obszarem badań w dziedzinie sztucznej inteligencji dopełniacza jest rozpoznawanie obrazów w sieciach neuronowych z wykorzystaniem sieci neuronowych (RNN) i jest to kolejne ważne podejście. Sieć ResNET jest wykorzystywana do umożliwienia poprawnego przetwarzania danych za pomocą rozpoznawania zadań i demonstracji za pomocą sieci neuronowych (DNN). Kolejne podejście tego systemu służy do przewidywania algorytmów, co pozwala neuronowym procesorom graficznym (GPU) wykorzystać paralelizm GPU. Podejście oparte na jednostkach przetwarzania grafów umożliwia wykorzystanie różnych modeli do analizy danych i znalezienia zadania dla konkretnego zestawu danych. Proponowany system oferuje różnorodne analizy, które służą do ulepszenia algorytmu sieci neuronowej w porównaniu z algorytmem konwencjonalnym. Ten system bada tłumaczenie maszynowe przy użyciu sieci neuronowej, która opiera się na mechanizmie rekurencyjnym z samouważnością. Jest on używany do redukcji złożoności systemu dzięki wydajnej architekturze neuronowej dla tłumaczenia maszynowego. Ta metoda jest używana do równoważenia złożoności liniowej w oparciu o dostrajanie zdań. W tradycyjnej metodzie występuje pewne opóźnienie w celu zrównoważenia sekwencji wejściowej w oparciu o liczbę tokenów wejściowych i złożoność nieliniową w przypadku transmisji długich sekwencji. Proponowany system [10] jest analizowany przy użyciu sieci uwagi strukturalnej. Ta sieć nie jest taka jak RNN, ale jest oparta na CNN pod względem liczby wejść i zaplanowanego zadania. System adresuje samouważność, która jest używana do analizy ograniczeń użytkownika w oparciu o jego potrzeby i dostępność zadań. Proces, taki jak po prostu wymusza sekwencyjne dane, jest używany do weryfikacji sekwencji i umożliwia modelowi ustrukturyzowanie uwagi, która jest wykorzystywana

do poprawy wydajności systemu. To podejście opiera się na osadzaniu danych, które są bardziej odpowiednie do reagowania na użytkownika za pomocą wielu systemów przewidywania zadań. Ten proces przechwytywania służy do poprawy jakości zdań i ułatwia przetwarzanie zadania w celu identyfikacji postępów i podobieństwa w tekście oraz postępów w reprezentowaniu zdań i uczeniu się. To podejście koncentruje się na poprawie jakości między maszyną a człowiekiem w oparciu o technologię tłumaczenia. Jakość tłumaczenia językowego jest bardzo ważna i jest osiągana poprzez wykorzystanie mechanizmu uwagi, danych szkoleniowych na dużą skalę i systemu Google NMT. Ta metoda wywiera wpływ na dziedzinę tłumaczenia maszynowego i adaptacji w sieciach neuronowych. W tradycyjnej metodzie NMT musi poprawić wydajność tłumaczenia, co jest przezwyciężane przez wprowadzenie tego alternatywnego podejścia poprzez połączenie wielu jednostek w jedną sieć. To ułatwia tłumaczenie danych wejściowych i wyjściowych za pomocą połączeń szybkiego przewijania. Istnieje bardziej zaawansowana technika tłumaczenia, która tworzy wydajną i skuteczną sieć. Stanford Alpaca ma strukturę, która rozumie i działa na polecenia werbalne, demonstrując swoją zdolność do wykonywania różnych czynności w oparciu o dane wejściowe pisemne. Dzięki wykorzystaniu obecnych postępów w rozumieniu języka naturalnego i uczeniu się przez wzmacnianie, model Alpaca wykazuje wyjątkową wydajność w przestrzeganiu instrukcji w różnych dziedzinach. Ta inicjatywa wnosi ważne pomysły, które można zastosować w rozwoju systemów AI, które mogą dokładnie rozumieć i wykonywać instrukcje ludzkie, a także są przydatne w robotyce, asystentach wirtualnych i systemach automatycznych. Przedstawiają nowe podejścia do zwiększania wydajności modeli rozumienia instrukcji wizualnych. W dostrajaniu instrukcji wizualnych autorzy sugerują , jak poprawić dopasowanie między instrukcjami wizualnymi a cechami wizualnymi poprzez jawne dostrajanie ich reprezentacji wizualnej za pomocą wyuczonej projekcji specyficznej dla instrukcji, umożliwiając lepsze rozumienie modelu i przestrzeganie takich poleceń w kontekście zadań wizualnych. Kolejna praca „Ulepszone linie bazowe z dostrajaniem instrukcji wizualnych" dalej udoskonala to podejście, pokazując jego skuteczność, o czym świadczy zwiększona wydajność w różnych testach porównawczych dotyczących rozumienia instrukcji wizualnych. Te prace razem przesuwają granicę wiedzy w zakresie rozumienia instrukcji wizualnych, proponując nowe techniki, aby wypełnić lukę między instrukcjami językowymi a percepcją wizualną w systemach AI. Celem tych badań jest zaproponowanie unikalnej metody nauczania modeli wizualnych poprzez instrukcje z języka naturalnego. Pomysł, który proponują tutaj autorzy, dotyczy tego, w jaki sposób duże zbiory danych mogą być wykorzystane do nauki reprezentacji wizualnych poprzez wstępne przetwarzanie dużej liczby obrazów i tekstu, a następnie alidację modelu względem zbioru danych, takiego jak zbiory danych podpisów obrazów. Poprzez trenowanie na tych połączonych danych, model uczy się kojarzyć cechy wizualne z odpowiadającymi im opisami tekstowymi, umożliwiając mu rozumienie i generowanie opisów obrazów w języku naturalnym. Praca ta znacząco rozwija dziedzinę uczenia multimodalnego, demonstrując skuteczność uczenia się reprezentacji wizualnych z nadzoru tekstowego i torując drogę do lepszego zrozumienia i generowania treści wizualnych przez systemy AI. Wynalezienie nowego sposobu podejścia do wstępnego trenowania multimediów wielordzeniowych. Jest on udoskonaleniem poprzedniego frameworka BLIP; w BLIP-2 wstępne trenowanie wiąże się z użyciem zamrożonych koderów obrazu i większych modeli językowych. Z punktu widzenia trenowania modelu, zamrożenie kodera obrazu z wyprzedzeniem oznacza, że możliwe jest wykorzystanie istniejących modeli wizji do destylacji wiedzy i poprawy wydajności w zadaniach downstream. Ponadto, włączenie dużych modeli językowych zwiększa bogactwo informacji kontekstowych, które mogą zostać przechwycone z danych tekstowych podczas wstępnego trenowania. Dlatego te aspekty sprawiają, że BLIP-2 jest ważniejszy w wielomodelowym wstępnym trenowaniu, ponieważ jest bardziej wydajny i skalowalny podczas uczenia się reprezentacji w modalnościach językowych i wizyjnych. Zapewnia nowe ramy do tworzenia modeli wizualnych o wysokiej wydajności. Teach BLIP rozszerza metodę BLIP (instrukcja przed instrukcją) poprzez włączenie konserwacji instrukcji, procesu, który zastępuje reprezentację modelu określonymi instrukcjami w

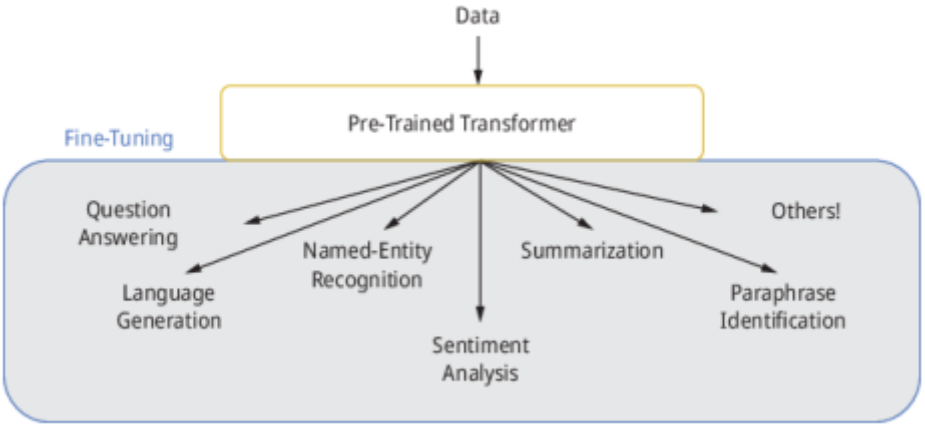
operacji. Poprawia to wydajność różnych interpretacji, umożliwiając modelowi lepsze zrozumienie i przestrzeganie specyfikacji. Wprowadzenie BLIP stanowi ważny krok w kierunku rozwoju ogólnego modelu języka, który można dostosować do różnych zadań i lekcji oraz przyczynia się do rozwoju różnych typów inteligencji, prezentując elastyczną inżynierię neuronową zdolną do obsługi zorganizowanych danych wejściowych i wyjściowych. Demonstracja Perceiver IO podejmuje wyzwanie przygotowania złożonych typów informacji, takich jak obrazy, dźwięki i treści, w sposób powiązany. Wykorzystując komponenty permutacyjno-ekwiwariantne i uczące się reprezentacje zbiorów, Perceiver IO może skutecznie obsługiwać zorganizowane informacje bez polegania na modelach specyficznych dla danego zadania. Ta adaptacyjność sprawia, że nadaje się on do szerokiego zakresu zadań, takich jak liczenie, klasyfikacja obrazów, wyszukiwanie pytań i obsługa dialektów. Perceiver IO reprezentuje znaczący postęp w planowaniu inżynierii neuronowej, promując powiązany system do obsługi zróżnicowanych modalności i zadań informacyjnych z ulepszoną sprawnością i wykonaniem. Zaproponowano nową metodę rozumienia wideo , która wykorzystuje wzorce językowe wyuczone na sprzęcie audiowizualnym, modyfikowane przez instrukcje. Video-LLAMA wykorzystuje informacje audio i wizualne oraz sugestie języka naturalnego, aby poprawić rozumienie treści wideo. Przetwarzając model jakości za pomocą konkretnych instrukcji, Video-LLAMA uczy się wyodrębniać istotne cechy z próbek audio i wizualnych, co pozwala mu na dokładną interpretację i analizę treści wideo. To podejście stanowi znaczący krok w kierunku bardziej intuicyjnych, bogatych w treść modeli rozumienia wideo, z zastosowaniami w takich obszarach jak podsumowanie wideo, rozpoznawanie akcji i rozumienie języka w kontekście. Zaplanuj sposób dostosowania podstawowego słownictwa do konkretnych zadań. Adaptacja niskiego rzędu (LoRA) zapewnia technikę ewaluacji niskiego poziomu, która zmniejsza obciążenie pracą związaną z dostrajaniem modeli językowych przed szkoleniem na dużą skalę, jednocześnie zachowując wydajność. Wykorzystując niskokosztowe szacunki, LoRA może szybciej dostosowywać się do nowych projektów, przy mniejszej liczbie zasobów pomiarowych i obliczeniowych. To podejście stanowi znaczący postęp w przenoszeniu modeli językowych do rzeczywistych aplikacji i oferuje rozwiązania na dużą skalę do wykorzystania wstępnie wytrenowanych modeli w różnych zadaniach, takich jak rozumienie i konstruowanie języka.

Dostrajanie LLM: Dostrajanie efektywne pod względem instrukcji i parametrów

LLM to ekspert w dziedzinie zaawansowanej akwizycji językowej (LLM), dysponujący silnymi umiejętnościami, takimi jak tworzenie tekstu, pisanie i rozumienie. Jednak ten LLM może oferować jedynie konkretne rozwiązania w określonych sytuacjach. Mogą istnieć zadania, których język potoczny nie byłby w stanie wykonać, w zależności od wybranej aplikacji. W takim przypadku jedną z możliwych alternatyw jest udoskonalenie LLM. Dostrajanie wymaga ponownego wytrenowania modelu bazowego na nowych danych. Chociaż optymalizacja jest zasobochłonna, trudna i niekoniecznie stanowi kluczowe rozwiązanie, jest ona bardzo skutecznym procesem i powinna być częścią zestawu narzędzi instytucji, które wdrażają LLM do swojej praktyki. Oto, co użytkownik powinien wiedzieć o dostrajaniu głównych modeli językowych. Nawet jeśli użytkownik nie posiada wystarczającej wiedzy, aby leczyć konkretnego użytkownika, znajomość sposobu działania pomoże mu szybko podjąć decyzję. Każdy model uczenia maszynowego można ulepszyć lub w niektórych przypadkach może wymagać modyfikacji. Celem trenowania modelu z wykorzystaniem zbiorów danych jest przewidywanie wzorców obecnych w samym rozkładzie danych. Rozważmy na przykład koncepcję CNN służącą do rozpoznawania obrazu samochodu. Model ten został wytrenowany na dziesiątkach tysięcy obrazów przedstawiających pasażerów w środowisku miejskim, co pokazano na rysunku



Są one powiązane z narzędziami i środowiskami, ponieważ ich trwałe ograniczenia dotyczą rozpoznawania obrazów, kolorów lub konfiguracji pikseli. Dzięki temu model ten może być z powodzeniem stosowany do obrazów pojazdów w krajobrazie miejskim. Sytuacja ulegnie jednak drastycznej zmianie, jeśli użytkownik spróbuje zastosować go do sterowania ciężarówkami na autostradzie. Nagle użytkownik zauważy, że wydajność modelu jest bardzo niska. Rozwiązaniem w tym przypadku jest trenowanie od podstaw, pracując szczególnie z obrazem ciężarówek na autostradzie. Wymaga to jednak stworzenia obszernej bazy danych zawierającej dziesiątki tysięcy obrazów tablic rejestracyjnych, co jest kosztowne i czasochłonne. Najlepsze wyniki przedstawiono na przykładzie klasyfikacji. Ciężarówki i autobusy mają wiele wspólnych cech. Niemniej jednak, aby to wdrożyć, konieczne jest stworzenie kompletnej bazy danych zawierającej dziesiątki tysięcy obrazów tablic rejestracyjnych, co jest kosztowne, długotrwałe i pracochłonne. Najlepsze wyniki przedstawiono w zależności od liczby przykładów klasyfikacji, na które można rozdzielić RA. Ponadto ciężarówki i autobusy stanowią dużą część tego samego kąta. Oznacza to, że można utworzyć wstępnie wytrenowany model, a nie od samego początku. Nawet przy niewielkiej bazie danych zdjęć pojazdów, a nawet dziesiątkach tysięcy lub tysiącach i kilku cyklach treningowych, użytkownik może dostosować istniejący model do nieznanego zadania. Zasadniczo dostrajanie odbywa się poprzez korygowanie niedopasowania modelu do rozkładu nowych danych. Obejmuje to koncepcję optymalizacji. Użytkownik bierze już wytrenowany model uczenia maszynowego i wykorzystuje nowe dane do udoskonalenia jego parametrów, dostosowuje go do nowego środowiska lub rozszerza do nowej aplikacji. Ta sama zasada dotyczy wzorców językowych. Jeśli model rozkładu danych treningowych użytkownika różni się od potrzeb innych aplikacji, dostrajanie może być dobrą decyzją. Na przykład, dobre CV może być wystarczające, jeśli użytkownik zatrudnia absolwenta studiów magisterskich (LLM) w praktyce medycznej, ale model edukacyjny musi zawierać dokumentację medyczną. Ważne jest jednak, aby dokładniej przeanalizować niespójności w jakości studiów magisterskich (LLM), ponieważ mają one specyficzne niuanse, które warto zbadać, co pokazano na rysunku 2.



Wybór wstępnie wytrenowanego modelu LLM

Ogólna analiza precyzyjnie dostrojonego modelu dużych zbiorów danych została przeprowadzona przy użyciu sieci GPT-3 i innych, które generują bardziej standardowe wyniki i wydajność, niezależnie od tekstu lub danych treningowych podanych przez użytkownika. Do każdej platformy AI, która jest używana do trenowania danych zgodnie z danymi wejściowymi zdefiniowanymi przez użytkownika i wytrenowanym modelem, przypisane są różne zadania treningowe języka naturalnego, a następnie wyświetlany jest wynik. Każdego dnia wykonano ponad 175 miliardów zadań z wykorzystaniem różnych funkcji LLM, co pokazano na rysunku 3.

Nawet użytkownik ma do dyspozycji model GPT-3 z wieloma ograniczeniami. Istnieje kilka darmowych modeli, takich jak BERT, które są dostępne dla dowolnej liczby zadań, dowolnego tekstu lub języka, wykorzystując wstępnie wytrenowany model LLM. Model ten służy do swobodnego dostrajania w oparciu o wytrenowany zbiór danych i analizuje różne podejścia, co daje efektywne wyniki dla różnych sekwencji. Użytkownik może przekazać dowolną formę danych, taką jak tekst, pytanie i odpowiedź w formie obrazu, w oparciu o model BERT, i zebrać efektywne odpowiedzi.

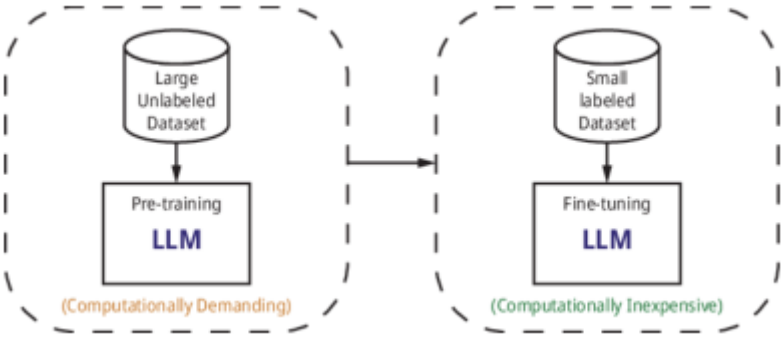
Różne podejścia do dostrajania modeli LLM

Ogólnie rzecz biorąc, podstawowa funkcja modelu LLM jest dostosowana do dowolnego typu danych, takich jak tekst, obraz, porównanie, przy użyciu różnych modeli LLM. Odbywa się to w oparciu o wstępnie wytrenowany model LLM oraz przy użyciu dowolnego mechanizmu nadzorującego i

wytrenowanych danych. To podejście jest wbudowane w dowolną funkcję LLM, co nadaje się do wszystkich zadań, od ogólnych po szczegółowe. System monitorowania LLM jest stabilny i nie wymaga aktualizacji, co oszczędza zasoby obliczeniowe. Aby jednak wytrenować klasyfikator, użytkownik musi posiadać bazę danych zawierającą tekst i odpowiadające mu kategorie w ramach krzywej uczenia. Zakres zbioru danych do precyzyjnego dostrajania będzie zależał od złożoności zadania i charakteru izolowanych obiektów. Jednak w niektórych przypadkach wartość wagi modelu transformatora może wymagać dostosowania. Aby to osiągnąć, użytkownik musi usunąć warstwę koloru i przeprowadzić proces precyzyjnego dostrajania w całym modelu. Praca ta może być obszerna i złożona, w zależności od rozmiaru modelu.

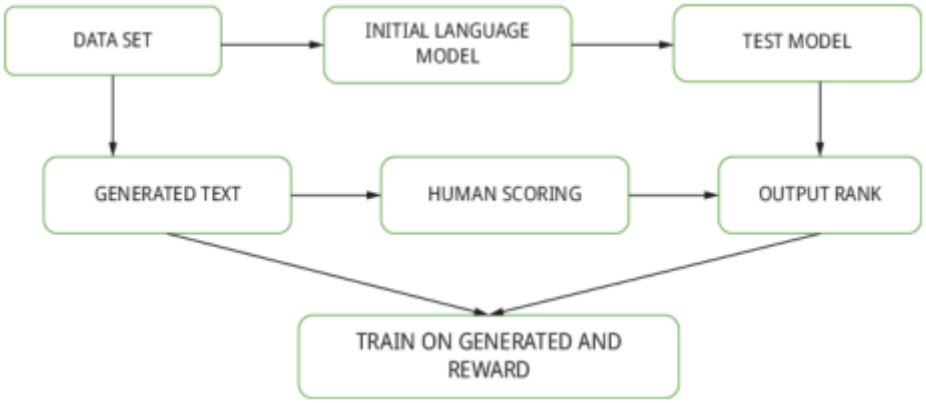
Dostrajanie nienadzorowane a dostrajanie nadzorowane (SFT)

W niektórych przypadkach może być konieczna zmiana bazy wiedzy bez zmiany działania programu LLM. Załóżmy, że użytkownik chce zaktualizować model na podstawie informacji medycznych lub nowego języka. W takich przypadkach przydatne mogą być informacje niestrukturyzowane, w tym artykuły w czasopismach medycznych i dokumentacjach akademickich, co pokazano na rysunku 6.4.



Uczenie się przez Wzmacnianie z Informacją Zwrotną (RLHF)

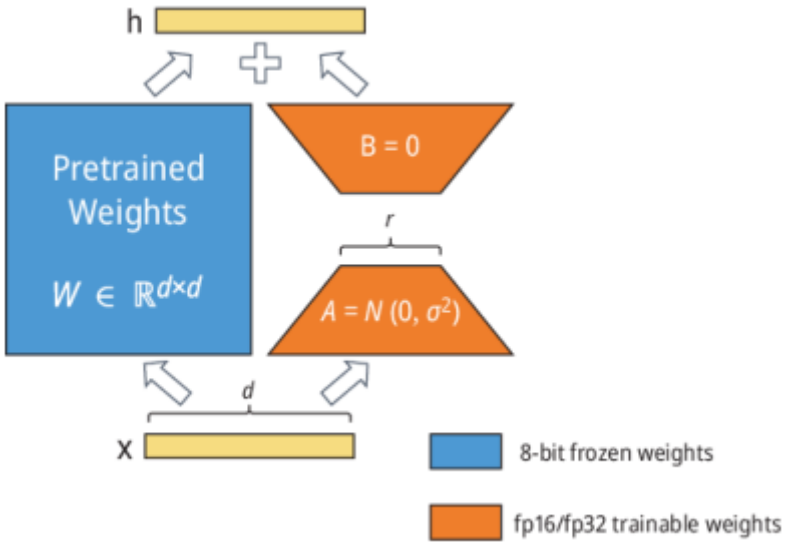
Niektóre organizacje nadal monitorują poprawę jakości lub szkolenia, aby wdrożyć wsparcie uczenia się poprzez informację zwrotną od zasobów ludzkich (RLHF). Ten złożony i efektywny proces wymaga wyboru audytorów i stworzenia modelu usług, który ułatwi doskonalenie LLM. Obecnie w RLHF mogą uczestniczyć tylko organizacje i agencje wywiadowcze dysponujące odpowiednimi talentami i możliwościami finansowymi. Dostępne są różne programy RLHF, ale najważniejszy z nich dotyczy rozwoju LLM. Po przeszkoleniu LLM z dużą liczbą tokenów, tworzona jest sekwencja tokenów, które najprawdopodobniej pojawią się w sekwencji. Tekst jest nadal spójny i zrozumiały, ale nie będzie on zrozumiały dla użytkowników ani aplikacji. RLHF oferuje analizę ludzi, aby pomóc LLM osiągnąć pożądane rezultaty. Audytorzy oceniają wyniki modelu pod kątem specyfikacji i w ten sposób rekomendują proces korekcji w celu osiągnięcia dobrych rezultatów, co przedstawiono na rysunku 5.



ChatGPT jest dobrym przykładem RLHF. OpenAI opracował ten model w oparciu o badania Instruct GPT. Początkowo przeprowadzono PFT, wykorzystując model GPT-3.5, instrukcje i odpowiedzi. Wybrano ewaluatorów, aby ocenić wyniki modelu w odniesieniu do różnych bodźców. Dane zwrotne od ludzi pomagają w opracowaniu modeli nagród, które symulują ludzkie preferencje. Modele językowe przechodzą następnie do głębokiego uczenia, gdzie tworzone są produkty, oceniany jest dla nich model nagród, a LLM optymalizuje ich parametry, aby zwiększyć ich rentowność.

Dostrajanie z Efektywnością Parametrów (PEFT)

Istotny obszar badań w zakresie dostrajania LLM koncentruje się na minimalizacji kosztów aktualizacji parametrów modelu. Dążenie to stanowi rdzeń dostrajania z Efektywnością Parametrów (PEFT), które obejmuje metody mające na celu redukcję liczby parametrów wymagających modyfikacji. Istnieje kilka podejść PEFT, a LoRA wyróżnia się jako wiodący przykład. LoRA zyskuje na popularności wśród modeli językowych o otwartym kodzie źródłowym. Zasada działania tej metody polega na tym, że dostrojenie modelu bazowego do konkretnych zadań często wymaga zmian jedynie w podzbiorze parametrów. W związku z tym macierz niskowymiarowa może trafnie uchwycić przestrzeń istotną dla zadania docelowego, pokazaną na rysunku 6.

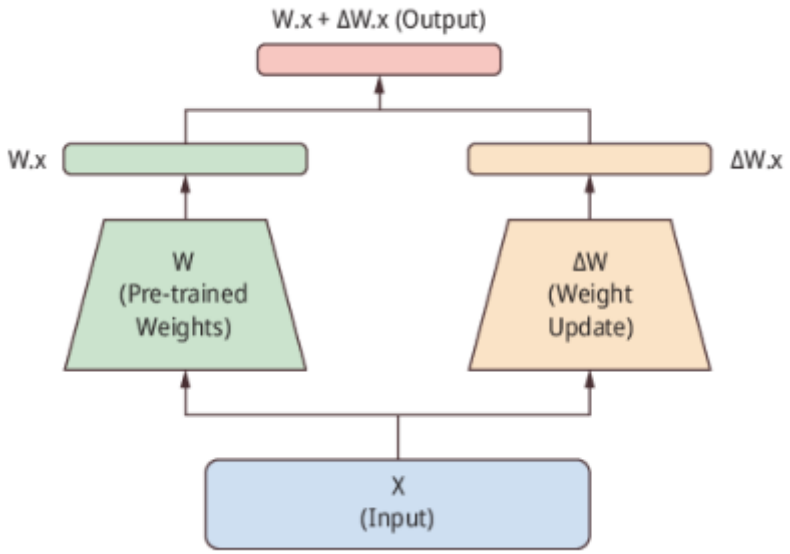


Wdrożenie podejścia LoRA polega na wykorzystaniu macierzy rang zredukowanych zamiast jednoczesnej modyfikacji parametrów podstawowego modelu LLM. Wartości parametrów z modelu LoRA są następnie scalane z głównym modelem LLM lub wykorzystywane podczas wnioskowania. Dzięki LoRA koszty dostrajania jakości można obniżyć nawet o 98%. Co więcej, umożliwia to

przechowywanie kilku kompaktowych, precyzyjnie dostrojonych modeli, które można łatwo włączyć do modelu LLM w czasie wykonywania. Dostrajanie dużych, wstępnie wytrenowanych modeli może być obliczeniowo dogłębne, często wymagając modyfikacji dziesiątek milionów parametrów. To tradycyjne podejście do precyzyjnego dostrajania, mimo że wydajne, wymaga znacznej mocy obliczeniowej i czasu, tworząc wąskie gardło dla unikalnych dla projektu wersji tych modeli. LoRA oferuje ekologiczne rozwiązanie poprzez rozbicie macierzy zmian podczas dostrajania jakości. Aby zagłębić się w LoRA, najpierw przyjrzyjmy się konwencjonalnemu systemowi dostrajania jakości.

– Dekompozycja ΔW

W konwencjonalnej metodzie dostrajania zadowalającego, użytkownicy mogą regulować wagi wstępnie wytrenowanej sieci neuronowej, aby dopasować ją do nowego zadania. Ta wersja zazwyczaj obejmuje modyfikację unikalnej macierzy wag (W) sieci. Zmiany wprowadzone do W w momencie dostrajania są oznaczane zbiorczo jako (ΔW), co skutkuje zaktualizowanymi wagami wyrażonymi jako (W Δ W). Jednak zamiast bezpośredniej edycji (W), metoda LoRA dąży do przerwania (ΔW). Ta dekompozycja ma kluczowe znaczenie dla zmniejszenia obciążenia obliczeniowego związanego z dostrajaniem wyjątkowo dużych modeli, jak pokazano na rysunku 7.



Hipoteza rangi wewnętrznej zakłada, że istotne zmiany w sieci neuronowej można przedstawić za pomocą formy o niższej liczbie wymiarów. Sugeruje ona, że nie wszystkie elementy (ΔW) mają równe znaczenie; raczej mniejszy podzbiór tych zmian może odpowiednio uchwycić wymagane korekty.

– Wprowadzenie macierzy (A) i (B)

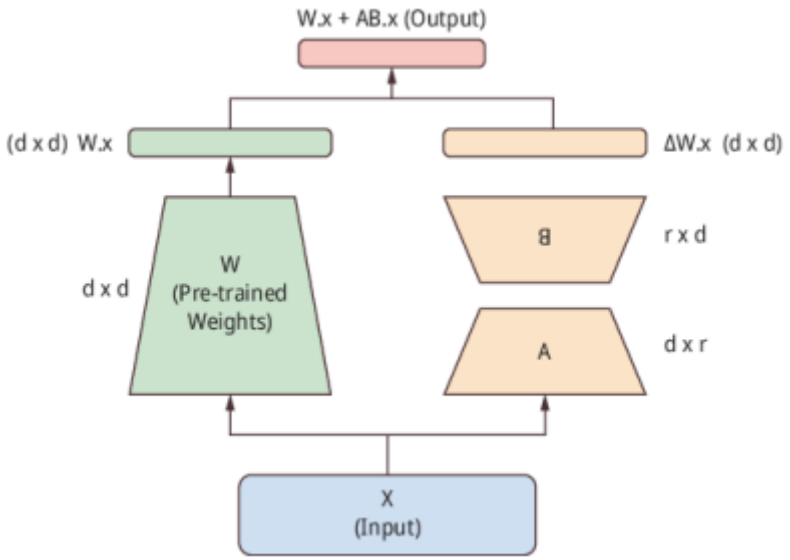
W oparciu o tę hipotezę, LoRA sugeruje wyrażenie (ΔW) jako wyniku mnożenia dwóch mniejszych macierzy, (A) i (B), z których obie mają obniżoną rangę. W związku z tym zaktualizowana macierz wag (W’) jest sformułowana jako

[W ’ = W + BA]

W tym równaniu W pozostaje zamrożone (tj. nie jest aktualizowane podczas treningu). Macierze (B) i (A) mają niższą wymiarowość, a ich iloczyn (BA) reprezentuje przybliżenie (ΔW) o niskim stopniu.

Wybierając macierze (A) i (B) o obniżonej randze (r), liczba parametrów możliwych do trenowania ulega znacznemu zmniejszeniu. Na przykład, jeśli (W) jest macierzą (d × d), konwencjonalna aktualizacja (W)

obejmowałaby parametry (d2). Jednak przy macierzach (B) i (A) o rozmiarach odpowiednio (d × r) i (r × d), liczba parametrów zmniejsza się do (2dr), znacznie mniejszej liczby, gdy (r ≪ d), jak pokazano na rysunku 8.



Zalety metody adaptacji niskiego rzędu (LoRA)

* Zmniejszone zużycie pamięci: LoRA zmniejsza zapotrzebowanie na pamięć poprzez minimalizację liczby parametrów wymagających aktualizacji, co pomaga w wydajniejszym przetwarzaniu modeli wielkoskalowych.

* Szybsze szkolenie i adaptacja: LoRA usprawnia potrzeby obliczeniowe, przyspieszając procesy szkolenia i dostrajania dużych modeli do nowych zadań.

* Zgodność ze sprzętem niższej klasy: Dzięki zmniejszonej liczbie parametrów, LoRA umożliwia dostrajanie dużych modeli na mniej wydajnym sprzęcie, takim jak skromne procesory GPU lub CPU.

* Rozkładając (ΔW) na iloczyn macierzy niższego rzędu, LoRA skutecznie osiąga równowagę między adaptacją dużych, wstępnie wytrenowanych modeli do nowych zadań a utrzymaniem wydajności obliczeniowej. Podstawowa koncepcja rangi wewnętrznej ma kluczowe znaczenie dla tej równowagi, zapewniając, że podstawowe możliwości uczenia się modelu pozostają nienaruszone przy jednoczesnym użyciu znacznie mniejszej liczby parametrów.

* LoRA i QLoRA: skuteczne metody precyzyjnego dostrajania modeli LLM

- Dostrajanie

- Parametryczne efektywne dostrajanie (PEFT)

- LoRA

- QLoRA

- Czterobitowa liczba zmiennoprzecinkowa (normalna)

- Kwantyzacja

- Problemy z wartościami odstającymi

- Kwantyzacja k-bitowa blokowa

- Podwójna kwantyzacja

- Dekwantyzacja

- Optymalizatory stronicowane

Metody PEFT

* LoRA

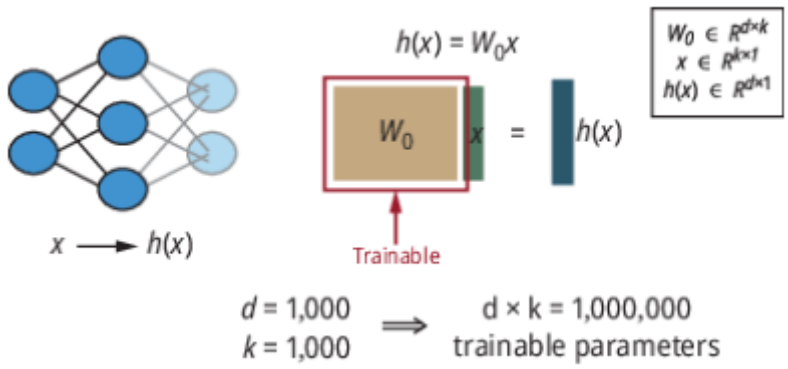
* Strojenie prefiksów

* Strojenie P

* Strojenie komunikatów

* QLoRA

W tym blogu system będzie koncentrował się wyłącznie na LoRA i QLoRA. Pozostałe metody zostaną omówione w kolejnych wpisach.



W tradycyjnym dostrajaniu modelu bez użycia PEFT, warstwa ukryta wykorzystuje macierz wag W0 z parametrami treningowymi (d × k), gdzie x ma wymiary (k × 1), a h(x) ma wymiary (d × 1). W rezultacie rozmiar parametru staje się znaczny, co komplikuje zarówno proces wstępnego treningu, jak i dostrajania, jak pokazano na rysunku 9.

LoRA: Metoda adaptacji niskiego rzędu

Jest to 16-bitowy transformator, który umożliwia precyzyjne dostrojenie poprzez dostosowanie tylko ograniczonego zestawu dodatkowych wag w modelu, przy jednoczesnym zachowaniu większości wstępnie wytrenowanych parametrów sieci. Zasadniczo użytkownik nie może modyfikować oryginalnych wag; zamiast tego wprowadza i trenuje dodatkowe wagi obok nich.

LoRA: Metoda adaptacji niskiego rzędu

Jest to 16-bitowy transformator, który umożliwia precyzyjne dostrajanie poprzez dostosowanie jedynie ograniczonego zestawu dodatkowych wag w modelu, przy jednoczesnym zachowaniu większości wstępnie wytrenowanych parametrów sieciowych. Zasadniczo użytkownik nie może modyfikować oryginalnych wag; zamiast tego wprowadza i trenuje dodatkowe wagi obok nich. W tradycyjnym podejściu dostrajania bez LoRA, warstwa ukryta używa wagi (W0 + ∇W). W tym przypadku oryginalna waga W0 pozostaje nienaruszona i nie jest wykorzystywana podczas precyzyjnego

dostrajania, skutecznie pozostając zamrożoną. Nowo wprowadzona waga, reprezentowana przez ∇W, jest uzyskiwana z iloczynu macierzy B i A, gdzie B ma wymiary (d × r), a A ma wymiary (r × k), jak pokazano na →rysunku 6.10. To r oznacza rangę macierzy aktualizacji i jest wyrażone jako liczba całkowita. Niższa ranga odpowiada mniejszym macierzom aktualizacji z mniejszą liczbą parametrów nadających się do trenowania. W tym kontekście wybraliśmy r równe 2, czyli mniejszą wartość, co skutkuje mniejszą liczbą parametrów nadających się do trenowania. W rezultacie mamy teraz tylko ((d + k) × r) parametrów nadających się do trenowania, czyli mniej niż pierwotnie (d × k) parametrów.

QLoRA: Kwantowana Metoda Adaptacji Niskiego Rzędu

To 4-bitowy transformator. QLoRA to technika precyzyjnego dostrajania, która łączy obliczenia o wysokiej precyzji z pamięcią masową o niskiej precyzji. Takie podejście pozwala zachować kompaktowy rozmiar modelu przy jednoczesnym zachowaniu wysokiej wydajności i dokładności. QLoRA wykorzystuje LoRA do precyzyjnego odwzorowania, mapując wartości 4-bitowe z powrotem do ich pierwotnego zakresu. Chociaż do obliczeń używane są zdekwantyzowane wagi, pozostają one przechowywane w pamięci w swojej 4-bitowej skwantyzowanej postaci. Podwójna dekwantyzacja: Odnosi się to do kwantyzacji stałych używanych w procesie kwantyzacji 4-bitowej funkcji NF. Ta pozornie niewielka korekta pozwala zaoszczędzić średnio 0,5 bita na parametr, jak opisano szczegółowo w powiązanych badaniach. Ta optymalizacja jest szczególnie skuteczna w kontekście QLoRA, która wykorzystuje blokową kwantyzację k-bitową. W przeciwieństwie do kwantyzacji wszystkich wag razem, ta metoda rozdziela wagi na oddzielne bloki lub fragmenty w celu niezależnej kwantyzacji .

Wnioski

Od adaptacji modeli do zupełnie nowych zobowiązań i radzenia sobie z rekordami spoza dystrybucji, po stosowanie nadzorowanego dostrajania jakości, niniejszy przewodnik zapewnił dogłębną analizę dostępnych technik. Ponadto, użytkownicy zagłębili się w uczenie wzmacniające z komentarzy ludzkich i dostrajanie jakości parametrów (PRE) jako zaawansowane techniki zwiększania możliwości LLM. Realistyczne studia przypadków podkreśliły namacalne korzyści płynące z dostrojonych programów LLM. Przykłady obejmują transformację analizy raportów karnych, pomoc w analizie nastrojów na rynkach finansowych, demonstrując rewolucyjny potencjał tej techniki w różnych sektorach. Podsumowując, dostrajanie jakości łączy zalety wstępnie wyszkolonych modeli ze specjalistyczną wiedzą, umożliwiając firmom pełne wykorzystanie przetwarzania języka naturalnego do realizacji ich ambitnych celów. W miarę postępu technologicznego, współpraca między programami LLM a ludzką kreatywnością PEFT obiecuje otworzyć nowe horyzonty innowacji, wydajności i percepcji w niezliczonej liczbie obszarów.