Model językowy do wytwarzania długich tekstów

Polscy badacze opracowali duży model językowy LongLLaMA, oparty na oprogramowaniu OpenLLaMA, stworzonym przez Meta. Jest on dostępny dla każdego w internecie.

Duże otwarte modele językowe o otwartym kodzie źródłowym pozwalają naukowcom na zaawansowane prace. Mogą być wykorzystywane do wszystkich zadań, w których ludziom już teraz pomagają chatboty. Chodzi np. o generowanie tekstu, edycję tekstu, rozmowę z użytkownikiem, tworzenie streszczeń czy tłumaczenie.

LongLLaMA w przeciwieństwie do ChatGPT nie posiada interfejsu w internecie, ale każdy może pobrać model ze strony HuggingFace i uruchomić go na własnym komputerze.

Model potencjalnie pozwoli obsługiwać 64 razy więcej tekstu niż ChatGPT - stwierdzają jego twórcy w informacji prasowej przesłanej PAP.

LongLLaMA opracowali: Szymon Tworkowski, Konrad Staniszewski, Mikołaj Pacek i Piotr Miłoś - badacze związani z IDEAS NCBR, Uniwersytetem Warszawskim i Polską Akademią Nauk, oraz Yuhuai Wu - jeden ze współtwórców xAI, startupu Elona Muska, i Henryk Michalewski - związany z UW i Google DeepMind.

"LongLLaMA to 'polski' duży model językowy, dostępny dla każdego w internecie. Może obsługiwać jednorazowo 8 tysięcy tokenów, czyli w przybliżeniu 30-50 stron tekstu, a w przypadku niektórych zadań znacznie więcej, nawet 256 tysięcy tokenów, chociaż to tylko wynik techniczny" – mówi lider zespołu dr hab. Piotr Miłoś.

Kiedy Meta, właściciel Facebooka, wypuściła OpenLLaMA, naukowcy z całego świata, między innymi pracujący pod kierunkiem prof. Miłosia, wzięli go na warsztat i modyfikowali.

"Nasza LongLLaMA jest w stanie przetwarzać znacznie większy kontekst niż było to wcześniej możliwe, czyli potrafi w jednym kawałku 'zjeść' znacznie więcej tekstu" - wyjaśnia prof. Miłoś.

Jak tłumaczy, LongLLaMA potrafi przetwarzać bardzo długie dane wejściowe. Dzięki temu generuje bardziej spójne i trafne odpowiedzi niż inne modele.

LongLLaMA może obsłużyć dowolną ilość kontekstu bez obcinania go i wypełniania, co pokazały testy z hasłem (passkey).

Badacze sprawdzali, czy po otrzymaniu bardzo długiego promptu (złożonego polecenia) LongLLaMA będzie w stanie przypomnieć sobie hasło podane na początku. OpenLLaMA dawała sobie radę tylko z promptem o długości 2 tysięcy tokenów, a przy dłuższych kontekstach jej efektywność spadała do zera. Natomiast LongLLaMA utrzymywała 94,5 proc. dokładności po otrzymaniu promptu o długości 100 tysięcy tokenów i 73 proc. dokładności po otrzymaniu 256 tysięcy tokenów.

Model ten potrafi obecnie wytwarzać spójne teksty o długości 8 tysięcy tokenów. Potencjalnie - nawet 256 tysięcy tokenów, w czym znacząco przewyższyłby m.in. ChatGPT - oceniają twórcy. Zużywa przy tym stosunkowo mało energii – do korzystania z LongLLaMA wystarczy pojedynczy procesor – i pracuje bardzo szybko.

"Jak wyobrazić sobie różnicę? Gdyby dla uproszczenia przyjąć, że 1 token to 1 słowo, podkreślmy, że 2 tysiące słów posiada mniej więcej 7-stronicowy artykuł. 256 tysięcy słów to w przybliżeniu długość powieści Harry Potter i Zakon Feniksa (257 tys. słów) albo Ulissesa (265 tys. słów)" - porównują polscy naukowcy.

"ChatGPT jest produktem komercyjnym. Został optymalizowany pod przyjemną obsługę. Modele takie jak LongLLaMA wydają raczej surowe informacje, na których dopiero można coś zbudować, np. analizować tekst albo produkować kod" – wyjaśnia prof. Miłoś.

Otwarte oprogramowanie mogą modyfikować informatycy na całym świecie, co odróżnia je od oprogramowania ChatGPT, które nie zostało udostępnione publicznie, choć wiadomo, że również bazuje na architekturze Transformer.

Jak wyjaśniają autorzy polskiego modelu, jest to rodzaj architektury sieci neuronowej, która analizuje tekst, aby rozróżnić skomplikowane powiązania między słowami na wielu warstwach, ucząc się wzorców na podstawie ogromnych ilości danych.

Technologia ta zrewolucjonizowała przetwarzanie języka naturalnego, umożliwiając chatbotom generowanie tekstu, tłumaczenie, rozmawianie z użytkownikiem i wiele innych zadań na poziomie niedostępnym wcześniej dla sztucznej inteligencji.

Prof. Miłoś tłumaczy, że kiedy zadajemy pytanie chatbotowi korzystającemu z Transformera, zmienia on tekst na tokeny. Są to fragmenty informacji, zwykle mające długość pomiędzy jednym znakiem a jednym słowem. W zdaniu „W 2023 roku, niespodziewanie, chatboty zmieniły nasze życie.” czatbot może zobaczyć przykładowo siedem słów, liczbę 2023, dwa przecinki i kropkę. Dzięki dzieleniu tekstu na tokeny sztuczna inteligencja potrafi efektywnie przetwarzać informacje.

Jednak liczba tokenów, jaką może przyjąć chatbot jest ograniczona – w przypadku ChatGPT 3.5 limit tokenów wynosi 4096, OpenLLaMA – 2000, a Google Bard – około 1000.

Dlatego, gdy zadajemy chatbotowi długie pytanie lub podajemy dużo informacji, może być konieczne ucięcie lub pominięcie niektórych fragmentów, aby zmieścić się w limicie tokenów. Większość istniejących chatbotów nie potrafi analizować całej książki, długiej rozmowy czy artykułu.

"Pełny potencjał dużych modeli językowych jest często ograniczony ze względu na to, ile kontekstu może przyjąć dany model – mówi Piotr Miłoś. – Dlatego wprowadziliśmy Focused Transformer (FoT), technikę wykorzystującą proces szkoleniowy inspirowany uczeniem kontrastowym (contrastive learning). To nowatorskie podejście pozwala na strojenie (fine-tuning) dostępnych już LLM, tak by były zdolne przyjmować większy kontekst".

Jak ocenia badacz IDEAS NCBR i PAN, LongLLaMA to duże osiągnięcie, ponieważ pokazuje, że duże modele językowe mogą pokonać ograniczenia związane z długością promptów i wytwarzać długie teksty, które będą przydatne dla człowieka.

Źródło: pap.pl
https://laboratoria.net/aktualnosci/31964.html

02-07-2026

Nośniki eków po 14 miesiącach na

Międzynarodowej Stacji Kosmicznej

Analizy mają pokazać, jak promieniowanie kosmiczne wpłynęło na nośniki leków.

23-06-2026

Flexicon FPC50 w dydaktyce pracy laboratoryjnej

Dostawca szkoleń aptaskil przygotowuje wykwalifikowanych specjalistów.

22-06-2026

Blisko 2,8 mln zł na badania nad terapią

Opracowanie strategii leczenia nowotworów odpornych na terapię.

22-06-2026

Studenci AGH zaprezentowali swój najnowszy bolid elektryczny

Pojazd powstał z myślą o udziale w zawodach inżyniersko-wyścigowych.

22-06-2026

Naukowcy sprawdzili, czy protony są wieczne

W badaniach uczestniczyły polskie ośrodki.

22-06-2026

Polska wśród krajów z najniższym poziomem stresu psychicznego

Wśród ukraińskich uchodźców.

22-06-2026

Życie seksualne coraz częściej przenosi się do świata technologii

Sfera ta rośnie szybciej niż wiedza o jej wpływie na ludzką seksualność.

22-06-2026

Przyjemnych snów życzy anestezjolog

Wystarczy przestrzegać protokołu znieczulenia.

Informacje dnia: Nośniki eków po 14 miesiącach na Międzynarodowej Stacji Kosmicznej Flexicon FPC50 w dydaktyce pracy laboratoryjnej Blisko 2,8 mln zł na badania nad terapią Studenci AGH zaprezentowali swój najnowszy bolid elektryczny Naukowcy sprawdzili, czy protony są wieczne Polska wśród krajów z najniższym poziomem stresu psychicznego Nośniki eków po 14 miesiącach na Międzynarodowej Stacji Kosmicznej Flexicon FPC50 w dydaktyce pracy laboratoryjnej Blisko 2,8 mln zł na badania nad terapią Studenci AGH zaprezentowali swój najnowszy bolid elektryczny Naukowcy sprawdzili, czy protony są wieczne Polska wśród krajów z najniższym poziomem stresu psychicznego Nośniki eków po 14 miesiącach na Międzynarodowej Stacji Kosmicznej Flexicon FPC50 w dydaktyce pracy laboratoryjnej Blisko 2,8 mln zł na badania nad terapią Studenci AGH zaprezentowali swój najnowszy bolid elektryczny Naukowcy sprawdzili, czy protony są wieczne Polska wśród krajów z najniższym poziomem stresu psychicznego

Partnerzy