AI je stále horké téma, ukázal Seznam Meetup věnovaný jazykovým modelům

29.10.2024 - Petra Barančíková | Seznam

17. října se uskutečnil už druhý ročník Seznam Meetupu zaměřeného na vývoj a praktické použití generativních jazykových modelů.

Akce přilákala pestrou směsici technologických nadšenců, výzkumníků i odborníků z oblasti umělé inteligence. Meetup byl součástí Dnů AI 2024 a přinesl řadu inspirativních přednášek, praktických ukázek a networkingových příležitostí. O účast na akci, která proběhla v budově ČVUT i online, projevilo zájem bezmála 600 účastníků.

O úvod meetupu se postaral tým ze Seznamu, hosty přivítali a celým programem provázeli Veronika Krejčířová a Vítek Líbal. Sérii přednášek odstartovala Diana Hlaváčová, která má v Seznamu produktově na starosti vývoj interního jazykového modelu – SeLLMy.

Diana Hlaváčová: Šelmy na scéně! Jak uvádíme SeLLMa modely do produkce

Diana se ve své přednášce zaměřila na klíčové důvody, proč Seznam investuje do vývoje vlastního velkého jazykového modelu. Zdůraznila především aspekty jako bezpečnost a efektivita. „Naším cílem je mít plnou kontrolu nad modelem, který běží na našich serverech, a zajistit, že uživatelská data zůstanou uvnitř firmy," uvedla.

Diana hovořila také o technických požadavcích na provoz více modelů současně, důrazu na nízkou latenci a efektivním využití hardwaru. Zmínila systém LLM proxy, který umožňuje snadnou adopci modelu díky jednotnému API využitelnému jak pro interní, tak externí aplikace. Prezentace zahrnovala i praktické ukázky aplikací, jako jsou sumarizace výsledků vyhledávání nebo generování popisků pro Sreality.

Jan Petrov: Od štěněte k SeLLMě: Co jsme před rokem nevěděli

Jan Petrov ze Seznamu představil technické aspekty fine-tuningu velkých jazykových modelů a jejich provoz na více grafických kartách. Zdůraznil, že u tak velkých modelů, jako je ten se 70 miliardami parametrů, je nutné používat technologie jako DeepSpeed Zero, která efektivně řídí distribuci dat během trénování. Zároveň zmínil důležitost syntetických dat a pokročilých metod, jako jsou rejection sampling a optimalizace modelů pomocí direct preference.

Honza se dotkl také problematiky dlouhého kontextu, který představuje výzvu pro modely s velkými nároky na paměť. V souvislosti s tím popsal, jak kvantizace a speciální techniky, jako je grouped query attention, mohou zlepšit efektivitu modelu při zpracovávání dlouhých textů.

Adam Kolář: Jazykové modely v klinických studiích

Zajímavým příkladem využití jazykových modelů v reálném světě byla přednáška Adama Koláře ze společnosti MindMed, který hostům představil projekt zaměřený na měření úzkosti pomocí LLM. Projekt, který vychází z klinických studií s pacienty, zahrnoval zpracování zvukových nahrávek a jejich hodnocení na Hamiltonově škále úzkosti. Adam vysvětlil, jak jejich model využívá transkripci pomocí technologie WhisperX a jak se snaží předejít problémům s halucinacemi, které mohou negativně ovlivnit výsledky.

Adam také ukázal, jak strukturovaný přístup ke tvorbě promptů a správná segmentace textu zlepšují přesnost hodnocení. Tento přístup společnosti MindMed umožnil dosáhnout vysoké shody s lidskými hodnotiteli a identifikovat oblasti, kde je možné se dále zlepšovat.

Petr Šimeček: Možnosti a úskalí využití velkých jazykových modelů ve světě mediálního monitoringu

Petr Šimeček ze společnosti Mediaboard představil využití jazykových modelů v oblasti zpracování mediálního obsahu a sumarizace článků. Zmínil výzvy spojené s halucinacemi modelů při sumarizaci textu a ukázal, jak tým z Mediaboardu využívá gpt4o-mini pro efektivní shrnutí a detekci sentimentu. Zároveň se dotkl problematiky embedování textů a benchmarků pro hodnocení kvality výsledků, přičemž vyzdvihl potřebu kvalitní češtiny a strukturovaných promptů pro lepší jazykovou plynulost. K tomu pak ukázal vlastní benchmark veřejně dostupný na Githubu.

Stanislav Fort: Adversariální útoky

Stanislav Fort z DeepMindu uzavřel meetup přednáškou o adversariálních útocích na neuronové sítě, které mohou zmanipulovat výstupy modelů. Ukázal, jak mohou specifické změny ve vstupních datech způsobit zmatení modelu a současně i své vlastní řešení, kdy model mimo jiné používá klasifikátory na každé vrstvě modelu a je o to robustnější. Prezentace zdůraznila důležitost bezpečnostních opatření v oblasti AI, a to zejména u autonomních systémů, jako jsou samořiditelná auta, kde by takové útoky mohly mít katastrofální následky.

Postery vzbudily zájem a podnítily diskuzi

Součástí meetupu byla také poster session, ve které se představily čtyři postery. Dva z nich představily nové benchmarky a evaluace českých generativních modelů:

- ? (autor: Adam Jirkovský)
- **BenCzechMark: Českocentrický víceúlohový a vícemetrikový benchmark pro jazykové modely s duelovým hodnotícím mechanismem** (autor: Martin Fajčík)

Zbylé dva postery pak ukázaly praktické použití generativních modelů v aplikacích Seznamu:

- **Klasifikace webových stránek za účelem určení zájmů uživatelů s využitím LLM** (autoři: Jaroslav Veverka, Andrej Vojtuš)
- **Automatická sumarizace podobných článků pomocí LLM** (autor: Michal Chudoba)

Chcete být součástí týmu, který se v Seznamu podílí na vývoji velkých jazykových modelů? Mrkněte na volné pozice, které právě obsazujeme.

Po oficiální části programu mezi účastníky probíhala neformální konverzace nad AI technologiemi a aktuálními výzvami v oboru. Networking pokračoval u sklenky vína, kde byla tématem nejen budoucnost velkých jazykových modelů, ale řešily se i praktické otázky jejich implementace. Atmosféra byla skvělá a diskuse tak živé, že pokračovaly i poté, co jsme opustili prostory ČVUT. Setkání ukázalo, jak důležitá je pro technologickou komunitu výměna zkušeností a jaké nadšení kolem AI technologií stále panuje. A to nás moc těší.

Za tým organizátorů ze Seznamu Petra Barančíková

http://blog.seznam.cz/2024/10/ai-je-stale-horke-tema-ukazal-seznam-meetup-venovany-jazykovym-modelum