Velké jazykové modely (LLM)

Velké jazykové modely (Large Language Models - LLM) jsou pokročilé algoritmy umělé inteligence (AI) založené na hlubokém učení (Deep Learning), které jsou schopny porozumět, generovat a manipulovat s lidským jazykem.

Tyto modely jsou trénovány na masivním množství textových dat (knihy, články, kód, internetové diskuze), díky čemuž se učí statistické vazby mezi slovy a dokáží predikovat, jaký text by měl následovat.

1. Architektura a princip fungování

Většina moderních LLM je postavena na architektuře **Transformer**, kterou v roce 2017 představil Google (paper „Attention Is All You Need“).

Vysvětlení o co jde v Attention Is All You Need

Dominantní modely sekvenční transdukce jsou založeny na komplexních rekurentních nebo konvolučních neuronových sítích v konfiguraci kodér-dekodér. Nejvýkonnější modely také propojují kodér a dekodér prostřednictvím mechanismu pozornosti. Navrhujeme novou jednoduchou síťovou architekturu, Transformer, založenou výhradně na mechanismech pozornosti, která zcela vylučuje rekurence a konvoluce. Experimenty se dvěma úkoly strojového překladu ukazují, že tyto modely jsou kvalitnější, lépe paralelizovatelné a vyžadují výrazně méně času na trénink. Náš model dosahuje skóre 28,4 BLEU v úkolu překladu z angličtiny do němčiny WMT 2014, čímž překonává dosavadní nejlepší výsledky, včetně ansámblů, o více než 2 BLEU. V překladové úloze z angličtiny do francouzštiny WMT 2014 dosahuje náš model nového špičkového skóre BLEU 41,8 po 3,5 dnech tréninku na osmi GPU, což je zlomek nákladů na trénink nejlepších modelů z literatury. Ukazujeme, že Transformer se dobře generalizuje na jiné úkoly, a to díky úspěšnému použití na analýzu anglických větných členů jak s velkým, tak s omezeným množstvím trénovacích dat.

- **Tokenizace:** Text není zpracováván jako celá slova, ale jako „tokeny“ (části slov, slabiky).
- **Attention Mechanism (Mechanismus pozornosti):** Umožňuje modelu vážit důležitost různých slov ve větě bez ohledu na jejich vzdálenost (např. pochopení kontextu zájmena na konci dlouhého odstavce).
- **Parametry:** „Neurony“ sítě. Čím více parametrů model má (miliardy až biliony), tím je obvykle schopnější, ale náročnější na hardware.

2. Využití LLM v praxi

LLM nejsou jen o chatování. V IT a byznysu mají široké uplatnění:

- **Generování kódu:** Psaní funkcí, refactoring, hledání bugů (např. GitHub Copilot).
- **Analýza a sumarizace:** Zpracování dlouhých dokumentů, extrakce klíčových informací.
- **Překlad:** Vysoce kvalitní kontextové překlady mezi jazyky.
- **Kreativní psaní:** Marketingové texty, e-maily, scénáře.

- **RAG (Retrieval-Augmented Generation):** Propojení LLM s firemní databází pro odpovídání na dotazy nad vlastními daty.

3. Přehled jednotlivých modelů (Inteligencí)

Trh s LLM se dělí na **uzavřené (proprietary)** modely, které běží na serverech poskytovatele, a **otevřené (open-weights/source)** modely, které lze provozovat lokálně.

A. OpenAI (Rodina GPT)

Průkopník moderní éry generativní AI.

- **GPT-3.5 Turbo:** Rychlý, levný model, který odstartoval mánii kolem ChatGPT. Dnes již zastaralý.
- **GPT-4:** Dlouho považován za krále LLM. Vynikající v logice, kódování a složitých instrukcích.
- **GPT-4o (Omni):** Multimodální model (text, audio, video v reálném čase). Rychlejší a levnější než GPT-4.
- **o1 (Strawberry):** Nová třída modelů zaměřená na **„reasoning“** (uvažování). Před odpovědí „přemýšlí“ (Chain of Thought), což ho činí excelentním v matematice a programování, ale pomalejším pro běžný chat.

B. Google (Rodina Gemini)

Google sjednotil své předchozí projekty (PaLM, LaMDA) pod značku Gemini. Jsou nativně multimodální.

- **Gemini Nano:** Nejmenší verze, určená pro běh přímo v mobilních telefonech (Android).
- **Gemini Flash:** Optimalizovaný pro rychlost a efektivitu, velká kontextová paměť (až 1M tokenů).
- **Gemini Pro:** Zlatý střed, hlavní konkurent GPT-4o.
- **Gemini Ultra:** Nejvýkonnější model pro nejnáročnější úlohy.

C. Anthropic (Rodina Claude)

Firma založená bývalými zaměstnanci OpenAI, zaměřuje se na bezpečnost a etiku („Constitutional AI“).

- **Claude 3 Haiku:** Extrémně rychlý a levný model, ideální pro čtení velkého množství dat.
- **Claude 3.5 Sonnet:** Aktuálně (2024/2025) často hodnocen jako nejlepší model na světě pro kódování a psaní, překonávající GPT-4o v nuancích.
- **Claude 3 Opus:** Původní vlajková loď, velmi silná v kreativním psaní.

D. Meta (Rodina Llama)

Meta (Facebook) razí cestu **Open Weights**. Modely dává k dispozici komunitě zdarma.

- **Llama 2:** Starší generace, která definovala standard pro open-source.
- **Llama 3 (8B, 70B, 405B):** Současná špička open-source.
  - *8B:* Lehký model, běží na běžných GPU.
  - *70B:* Výkonný model srovnatelný s GPT-3.5/4.
  - *405B:* Masivní model konkurující GPT-4o, ale vyžaduje obrovský hardware.

E. Mistral AI (Evropská špička)

Francouzský startup, který je velmi efektivní a populární mezi vývojáři.

- **Mistral 7B:** Malý, ale velmi schopný model.
- **Mixtral 8x7B (MoE):** Využívá architekturu **Mixture of Experts**. Model se skládá z několika menších sítí, které se aktivují podle potřeby. Velmi rychlý a efektivní.
- **Mistral Large:** Uzavřený model, konkuruje GPT-4.
- **Codestral:** Specializovaný model pro programování.

F. Ostatní významné modely

- **Grok (xAI):** Model Elona Muska, integrovaný do sítě X (Twitter). Má přístup k reálným datům z této sítě a vyznačuje se menšími zábranami („vzpurný mód“).
- **Phi (Microsoft):** Série „Small Language Models“ (SLM). Trénované na učebnicových datech, aby byly extrémně malé, ale logicky zdatné.
- **Command R+ (Cohere):** Specialista na RAG a práci ve firemním prostředí, exceluje v citování zdrojů.

4. Výzvy a rizika

- **Halucinace:** LLM neumí „fakta“, pouze predikuje slova. Může sebevědomě tvrdit naprosté nesmysly.
- **Context Window (Kontextové okno):** Omezená paměť modelu. Jakmile konverzace přesáhne limit (např. 128k tokenů), model zapomíná začátek.
- **Bias (Předpojatost):** Modely přejímají stereotypy z trénovacích dat.

Tagy: ai llm gpt claude llama machine_learning