Čo chýba ChatGPT k tomu, aby rozumel, čo robí?

Martin Takáč

Centrum pre kognitívnu vedu FMFI UK
Mlynská dolina, 848 48 Bratislava
Email: martin.takac@fmph.uniba.sk

Abstrakt

Pre ChatGPT platí problém ukotvenia symbolov a všetky s ním spojené námietky. Napriek tomu nás ohromuje kvalitou vygenerovaných odpovedí a máme tendenciu veriť, že rozumie tomu, čo hovorí. Ale je to naozaj tak? V príspevku predstavím niekoľko pokusov ukotviť symboly veľkých jazykových modelov v multimodálnych dátach, resp. ich prepojiť s telom. Na záver naznačím disruptívne zmeny, ktoré systémy ako ChatGPT môžu spôsobiť, resp. už spôsobujú.

1 Úvod

V novembri 2022 dala spoločnosť OpenAI k verejnému používaniu systém ChatGPT. Ten okamžite pritiahol pozornosť médií aj používateľov – po prvých piatich dňoch ich bolo milión, v januári 2023 sto miliónov a dnes ich s ním denne interaguje približne 25 miliónov.[1] Dokáže konverzovať, písať básne aj eseje, riešiť matematické problémy, školské úlohy a testy, aj generovať kód v rôznych programovacích jazykoch. Súčasná (výrazne vylepšená) verzia s názvom GPT-4 sa v mnohých benchmarkových úlohách približuje k ľudskému výkonu (OpenAI, 2023). Výskumníci z Microsoftu vo svojom rozsiahlom reporte (Bubeck a spol., 2023) tvrdia, že GPT-4 možno považovať za prvotnú (zatiaľ neúplnú) verziu všeobecnej umelej inteligencie (AGI), teda umelej inteligencie, ktorá bez špeciálneho dotrénovania dokáže riešiť hocijakú úlohu, ktorú by vyriešil človek.

2 Ako to funguje

ChatGPT je veľký jazykový model (vo verzii GPT-4 obohatený o vizuálnu modalitu)—hlboká neurónová sieť s 10^{12} parametrami na báze transformera (Vaswani a spol., 2017), ktorý pre textový prompt na vstupe vygeneruje jeho pokračovanie (odpoveď). Odpovede hodnotené ľuďmi tvoria (spolu s promptom) trénovacie dáta pre reinforcement learning (RL) modul, ktorý sa naučí zoraďovať odpovede podľa vhodnosti.

[1] https://nerdynav.com/chatgpt-statistics/

3 Čo znamená „rozumieť“?

Umelé systémy sú tradične kritizované pre nedostatočné ukotvenie symbolov (Harnad, 1990): napriek tomu, že takýto systém môže správne reťaziť symboly, či odpovedať na otázku, pre neho sú ukotvené iba vo vzťahu k iným symbolom, ale nie k reálnemu svetu. Searle (1980) to opisuje metaforou Čínskej izby: Searle je zavretý v izbe, dostáva zvonka na papieri otázky v čínštine a pomocou veľkej knihy s pravidlami a porovnávania tvarov znakov podáva von na papieri správne odpovede v čínštine bez toho, aby rozumel po čínsky. Na argument Čínskej izby možno odpovedať viacerými spôsobmi (Cole, 2020). Jeden z nich je „The other minds reply“: o tom, že nám rozumejú iní ľudia, sa vieme presvedčiť iba pragmaticky, na základe ich správania. Ak správne zareagujú na našu požiadavku, tak jej rozumejú. Tento princíp tvorí základ mnohých variantov Turingovho testu: pokiaľ umelý systém v nejakej zložitej jazykovej úlohe uspeje ako človek, tak rozumie jazyku. GPT-4 by v mnohých ohľadoch Turingovým testom prešiel. Napriek tomu mu chýba ukotvenie symbolov a modely jeho typu boli označené za „stochastické papagáje“ (Bender a spol., 2021). Tento nedostatok sa moderné systémy snažia vyriešiť prepojením jazykovej domény s inými modalitami, najmä vizuálnou.

4 Multimodálne ukotvenie symbolov

Štandardným SOTA systémom sa stal CLIP (Radford a spol., 2021), ktorý prepája transformerový enkóder pre text s enkóderom pre obrázky. Je trénovaný metódou kontrastívneho učenia na pároch obrázok-text verejne dostupných z internetu. Oba enkódery sa trénujú na predikciu toho, ktoré obrázky boli spárované s akým textom. Následne sa system používa ako zero-shot klasifikátor, ktorý vie k obrázkom doplňať popisy.

V čase písania tohto článku bol predstavený multimodálny systém Kosmos-1 od Microsoftu (Huang a spol., 2023) zvládajúci tvorbu titulkov k obrázkom, odpovedanie na otázky o obrázkoch (VQA), rozpoznávanie entít v obrázkoch na základe textových inštrukcií, a ďalšie úlohy. V Ravenovom IQ teste (merajúcom schopnosť neverbálneho uvažovania) dosiahol 22 % úspešnosť bez dotrénovania oproti 17 % base-

line pri náhodnom hádaní. Systém pozostáva z transformerového dekódera trénovaného na obrovskom korpuse multimodálnych sekvencií (so špeciálnymi tagmi pre netextové data).

5 Prepojenie s telom

Ďalším (logickým) trendom je prepájanie jazykových modelov s robotickými telami. Microsoft (Vemprala a spol., 2023) reportuje o experimentoch, v ktorých dostal ChatGPT API pre detekciou objektov a vzdialeností a ovládanie robota, a textové popisy obrazu z kamery v každom časovom kroku. ChatGPT dokázal úspešne riadiť manipuláciu s objektami robotickým ramenom i ovládať dron.

Rovnakým smerom sa uberá aj výskum v Google. Článok (Driess a spol., 2023) predstavuje systém PaLM-E, ktorý pozostáva z dekodéra s 562 miliardami parametrov, na vstupe má multimodálne vety obsahujúce vizuálnu i textovú informáciu ako aj kontinuálne odhadovanie stavu systému. Je trénovaný end-to-end a zvláda VQA, titulkovanie obrázkov i riadenie robota.

6 Čínska izba ešte raz

Súčasné trendy v ukotvovaní jazykových modelov kopírujú princípy z niekoľkých ďalších odpovedí na argument čínskej izby. Podľa „System reply“ netreba postulovať v izbe Searla (rovnako ako v ľudskom mozgu sa nenachádza žiaden homunkulus), ale je to celý systém (izba, vstupy a výstupy, program a proces jeho vykonávania), ktorý rozumie čínsky. Podľa „Robot reply“ možno myšlienkový experiment pozmeniť tak, že Searle nebude v izbe, ale v robotovi, pričom vstupy prichádzajú z kamery a Searlove výstupy riadia efektory robota. A napokon, podľa „Developmental reply“ by v tomto robotovi nemusel byť dospelý Searle, ale novorodenec, ktorý si postupne osvojí a naučí sa všetky zákonitosti prepojení medzi vstupmi a výstupmi.

Ak teda dáme umelému systému telo a prepojíme jeho vstupy a výstupy na reálne prostredie a vybavíme ho silnými štatistickými mechanizmami učenia (napr. v hlbokej neurónovej sieti),[2] bude sa takýto systém principiálne líšiť od toho, ako funguje človek, jeho telo a jeho učiaci sa mozog?

7 Záver

V tomto článku som sa venoval najmä teoretickej otázke, v akom zmysle môžu systémy ako GPT-4 rozumieť tomu, čo robia. Tieto systémy so sebou však prinášajú aj množstvo urgentných praktických otázok. Doteraz bola kritizovaná najmä ich zaujatosť (biases), netransparentnosť a ekologické dopady (Bender a spol., 2021). Tým, že sa poskytli verejnosti v masovom meradle, sa stali disruptívnou technológiou, ktorá zmení obchodné modely fungovania, žurnalistiku, trh práce, vzdelávanie, nehovoriac o možnej destabilizácii demokracie, ak sa stanú nástrojom propagandy. Preto by bolo vhodné časť zdrojov, ktoré veľké firmy venujú na čo najrýchlejšie vytvorenie čo najvýkonnejšieho modelu, presmerovať do výskumu bezpečnostných aspektov takýchto modelov a ich dopadu na spoločnosť.

[2] Podstatným je zaväzbenie systému s prostredím v reálnom čase a rýchle (1-shot, few shot) a priebežné učenie, inak by vedomosti systému ostali zamrznuté v čase trénovania.

Poďakovanie

Vznik tohto článku bol podporený grantom VEGA 1/0373/23 a KEGA 022UK-4/2023.

Literatúra

Bender, E. M. a spol. (2021). On the dangers of stochastic parrots: Can language models be too big? V *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, str. 610–623.

Bubeck, S. a spol. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv:2303.12712 [cs.CL].

Cole, D. (2020). The Chinese Room Argument. Zalta, E. N. (zost.), V *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

Driess, D. a spol. (2023). PaLM-E: An embodied multimodal language model. arXiv:2303.03378 [cs.LG].

Harnad, S. (1990). The symbol grounding problem. *Physica D*, 42:335–346.

Huang, S. a spol. (2023). Language is not all you need: Aligning perception with language models. arXiv:2302.14045 [cs.CL].

OpenAI (2023). GPT-4 technical report. arXiv:2303.08774 [cs.CL].

Radford, A. a spol. (2021). Learning transferable visual models from natural language supervision. arXiv:2103.00020 [cs.CV].

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–457.

Vaswani, A. a spol. (2017). Attention is all you need. arXiv:1706.03762 [cs.CL].

Vemprala, S. a spol. (2023). ChatGPT for robotics: Design principles and model abilities. Technická správa MSR-TR-2023-8, Microsoft.