ALTERNATOR
Misliti znanost.

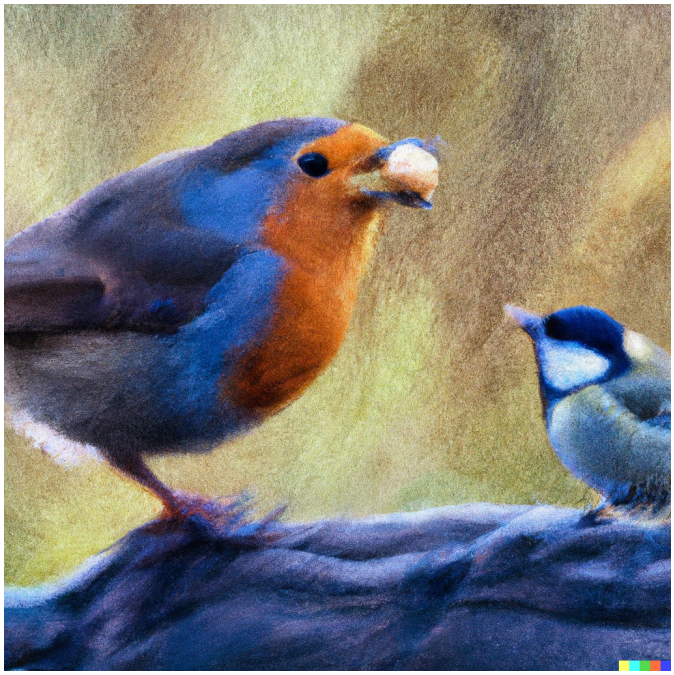
Veliki jezikovni velikani: so tudi prijazni?

6. 4. 2023

Številka: 14/2023

Avtor:

- Marko Robnik-Šikonja



Sliko je zgeneriral spletni servis DALL-E 2 (https://openai.com/product/dall-e-2) z ukazom »a giant robin feeding a tiny titmouse, impressionist style« (velikanska taščica hrani drobno siničko, impresionistični slog)

V zadnjih letih v medijih pogosto zasledimo novice in komentarje s področja umetne inteligence. Novice so večinoma dobre, govorijo o novih zmožnostih računalnikov, boljših digitalnih storitvah in novih uporabnih tehnologijah na različnih področjih. V nasprotju s tem so komentarji nemalokrat zaskrbljeni, govorijo o zatonu človeštva, tehnološki singularnosti, izgubi služb in možnih zlorabah. Ker ljudje načeloma močneje reagiramo na potencialne nevarnosti kot na morebitne ugodnosti, je rezultat strah, zaskrbljenost in vznemirjenje. Najboljše zdravilo za to je luč znanja, ki razblini temne sence in posveti v kotičke, kjer se skrivajo strahovi. Tokrat osvetljujemo velike jezikovne modele, kot je ChatGPT, ki trenutno poplavlja medijsko krajino.

Globoke nevronske mreže

Novejši uspehi interdisciplinarnega znanstvenega in tehnološkega področja umetne inteligence temeljijo predvsem na globokih nevronskih mrežah. Umetne nevronske mreže delujejo nekoliko podobno možganom. Podatke na vhodu skozi zaporedje slojev postopoma pretvorijo v bolj abstraktno in informativno predstavitev ter v njej poiščejo koristne vzorce. Da bi nevronsko mrežo česa naučili, je potrebna velika množica rešenih primerov nekega problema. Denimo, da imamo zbirko slik tkiv, ki so označene z diagnozo rak ali ne-rak. Vsako sliko iz zbirke pošljemo skozi mrežo, da napove verjetnost rakavosti tkiva. Izračunamo razliko med napovedmi in pravilnimi rezultati ter popravimo uteži na povezavah med nevroni tako, da mreža vrača pravilne rezultate. Proces popravljanja uteži mnogokrat ponovimo, dokler mreža pravilno ne napove večine slik. Naučena umetna nevronska mreža bo na novih slikah zmožna prepoznati rakavo tkivo.

Umetne nevronske mreže so znane že od leta 1943 in so skupaj s celotnim področjem umetne inteligence doživele že več vzponov in padcev, ki jih imenujemo zime umetne inteligence (https://en.wikipedia.org/wiki/AI_winter) . Daljše zime v letih 1974–1980 in 1987–1993 so bile posledice (pre)velikega optimizma in (pre)velikih obljub vodilnih raziskovalcev, ki so obljubljali skorajšnje velike preboje. Ker se ti niso realizirali, se je navdušenje javnosti in financerjev ohladilo in področje je dlje časa stagniralo. Nevronske mreže novo pomlad in veliko praktično uporabnost doživljajo od leta 2009 naprej z uporabo tisočih procesorjev na grafičnih karticah za njihovo hitro učenje. Istočasno s tem so se pojavile tudi velike zbirke označenih podatkov, npr. velike množice slik in besedil na internetu in Wikipediji. To omogoča strojno prepoznavanje objektov na slikah, razpoznavanje govora, strojno prevajanje, učinkovito iskanje po besedilih, odgovarjanje na vprašanja, povzemanje besedil in tudi klepetanje z umetnimi sistemi. Nova zima zaenkrat še ne prihaja.

V nasprotju z naravnimi nevroni evolucija njihovih umetnih soimenjakov poteka mnogo hitreje. Leta 2017 se je pojavila arhitektura nevronskih mrež, imenovana transformer (https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf) , ki je integrirala vrsto dobrih lastnosti predhodnih mrež: prilagojena je vzporednemu računanju na tisočih jedrih grafičnih

procesorjev, sprejme dokaj dolge vhode, za vsak vhod učinkovito upošteva kontekst ostalih vhodov, dobro pa se obnaša tudi pri velikih konfiguracijah z milijoni, milijardami in celo bilijardami parametrov. Nevronske mreže tipa transformer so bile prvotno razvite za strojno prevajanje, a so danes prevzele primat na celotnem področju obdelave naravnega jezika pa tudi na področju obdelave slik, videa, časovnih vrst in tudi mešanih jezikovno-slikovnih podatkov.

Veliki jezikovni modeli

Tudi jezikovni modeli so že stara pogruntavščina. V osnovi so namenjeni verjetnostnemu modeliranju jezika in tipično napovedujejo naslednjo ali predhodno besedo ali znak. V začetku so temeljili predvsem na statistikah skupin besed, npr. zaporedju dveh ali treh besed, izračunanih na velikih besedilnih zbirkah. Iz njih se že da napovedati naslednjo besedo, kar je koristno npr. pri prepoznavanju govora. Z nevronskimi mrežami so jezikovni modeli pridobili mnogo daljši kontekst, večjo sposobnost predstavitve besedil in generiranja človeku všečnih tekstov. Učijo se kar iz velikanskih množic besedil na internetu. Učenje poteka z napovedovanjem naslednje besede, pravilni odgovori pa so znani, saj jih lahko vzamemo iz dejanskih besedil.

Prvi velik jezikovni model arhitekture transformer z imenom BERT (http://dx.doi.org/10.18653/v1/N19-1423) so ustvarili Googlovi raziskovalci konec leta 2018. Namenjen je bil kontekstualni predstavitvi besedil, kar pomeni, da se besedilo iz simbolične oblike pretvori v številske vektorje, ki semantične lastnosti besedila pretvorijo v razdalje med vektorji. Takšna vektorska predstavitev besedila, ki jo imenujemo vložitev, omogoča vrsto računskih operacij, ki posnemajo človekovo razumevanje jezika. Model BERT se uči napovedovati skrite besede v stavkih, za kar skrijemo manjši del, npr. 15 % besed v realnem besedilu. V najmanjši različici je imel model BERT 110 milijonov parametrov, učil pa se je iz besedila celotne angleške Wikipedije (2,5 milijarde besed) in zbirke knjig (800 milijonov besed). Uspeh tega modela je temeljito spremenil področje obdelave jezika. Od leta 2019 naprej Googlov iskalnik vsako iskanje obdela z modelom tipa BERT, večina današnjih raziskav na področju naravnega jezika pa uporablja eno od inačic jezikovnih modelov arhitekture transformer.

Razvoj jezikovnih modelov je od leta 2019 naprej skokovit. Pojavile so se nove tehnološke inačice jezikovnih modelov ter jezikovne in tematske prilagoditve modela BERT. Modeli vrste T5 (*Text To Text Transfer Transformer* (https://www.jmlr.org/papers/v21/20-074.html)) so namenjeni poljubnim preslikavam med besedili, npr. poenostavljanju besedil ali pretvorbi v pravilno slovnično obliko. Generatorji besedil z imenom GPT (*Generative Pretrained Transformer*) so namenjeni različnim generativnim nalogam. Kot izhodišče vzamejo neko podano besedilo, ki služi kot kontekst, da na njegovi podlagi model izračuna verjetnostno porazdelitev naslednje besede. Z vzorčenjem iz porazdelitve model izbere besedo, jo generira in doda h kontekstu. Na podlagi novega konteksta spet generira naslednjo besedo. Postopek se nadaljuje, dokler se ne generira posebna beseda, ki označuje konec besedila. Model GPT lahko uporabimo za različne naloge, npr. pri povzemanju besedil model naučimo, da iz konteksta, ki predstavlja daljše besedilo, generira njegov povzetek. Pri odgovarjanju na vprašanja lahko kontekst predstavlja neko besedilo, iz katerega želimo pridobiti odgovore. Še splošnejše je odgovarjanje na poljubna vprašanja, kjer predvidevamo, da je model odgovor videl med postopkom vnaprejšnjega učenja, kontekst pa je kar vprašanje samo.

Model tretje generacija te arhitekture, GPT-3 (https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf) iz leta 2020, ima 175 milijard parametrov (človeški možgani denimo še vedno vsebujejo približno 1.000 krat več povezav med nevroni), naučen je na 500 milijardah besed, njegovo učenje pa je zahtevalo najem računske moči v ocenjeni vrednosti 12 milijonov dolarjev. Danes največji jezikovni model GLaM (https://proceedings.mlr.press/v162/du22c.html) vsebuje več kot 1,2 bilijardi parametrov. Žal učenje velikih jezikovnih modelov zahteva tudi porabo precejšnje količine električne energije, ki povzroča znatne izpuste ogljikovega dioksida – tudi okoljski vidiki velikih modelov so danes pod drobnogledom raziskovalcev in javnosti. Podobni veliki modeli arhitekture transformer so nastali tudi za področje obdelave slik in kombinacije slik in besedil.

Pogovorni jezikovni modeli

Generativni jezikovni model tipa GPT-3 so konec leta 2022 zelo uspešno prilagodili v pogovorni model ChatGPT (https://chat-gpt.org/) . Model je vzbudil izjemno zanimanje javnosti, saj se je izkazal kot odličen sogovornik z obširnim znanjem, presenetljivo (ne pa popolno) sposobnostjo sklepanja in gladkimi ter lingvistično pravilno generiranimi odgovori. ChatGPT je model vrste GPT-3 dopolnil z učenjem na učni množici človeških dialogov. Odgovore modela so uredniki popravili. Glede na uspešnost in pravilnost so rangirali kakovost več popravljenih odgovorov (https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf) in tako pridobili nov model, ki je ocenjeval kakovost prvotnega modela. Po nekaj iteracijah izboljševanja je bil pogovorni model ChatGPT pripravljen. Pri uporabi ima model vgrajene še dodatne kontrole, ki se nanašajo na neprimerna vprašanja, neprimerno besedišče (npr. seksistično ali rasistično) in področja, ki jih lastnik modela, podjetje OpenAI, noče pokrivati (npr. politično občutljive teme). Primerov uspešne rabe ne bomo prikazovali, raje delovanje preizkusite sami.

ChatGPT pozna številne jezike, vključno s slovenščino, in je zmožen reševati številne generativne probleme jezikovnega procesiranja, kot je odgovarjanje na vprašanja, povzemanje in poenostavljanje besedil. Dokaj uspešno na podlagi navodil generira tudi kodo v več programskih jezikih.

V marcu 2023 se je pojavil model GPT-4 (https://doi.org/10.48550/arXiv.2303.08774) , ki je še izboljšana verzija modela ChatGPT. V polni komercialni verziji ta model omogoča vstavljanje slik in izboljšuje nekatere druge lastnosti modela

ChatGPT. Konkurenca ne miruje in v uporabi ali najavljeni so še drugi pogovorni jezikovni modeli, naj omenimo le zanimive novosti iskalnika and klepetalnika you.com (https://you.com/) . Tudi v EU se obeta izgradnja zelo velikih prostodostopnih jezikovnih modelov, primerljivih z GPT-3 in ChatGPT. Slovenska jezikovnotehnološka skupnost je vključena v francosko in nemško pobudo na tem področju.

Zmožnosti novih velikih jezikovnih modelov zbujajo tako navdušenje kot zaskrbljenost. Nekateri zaskrbljeni posamezniki so celo pripravili poziv, da se za nekaj časa zaustavi gradnja (https://futureoflife.org/open-letter/pause-giant-ai-experiments/) jezikovnih modelov, boljših od GPT-4. Temu drugi nasprotujejo z argumenti (https://aisnakeoil.substack.com/p/a-misleading-open-letter-about-sci?publication_id=1008003&post_id=111502892&isFreemail=true) , da kritike navajajo špekulativna futuristična tveganja, ne naslovijo pa dejanskih težav z napačnimi informacijami, vplivom na delovna mesta in varnostjo.

Omejitve velikih jezikovnih modelov

Z znanstvenega vidika je precejšnja omejitev modela ChatGPT njegova zaprtost, saj ni povsem jasno, na katerih podatkih je bil naučen. Podatki modela časovno segajo do septembra 2021 in model kasnejših dogodkov ne pozna. Model se ne uči na podlagi lastnih izkušenj in ne išče odgovorov v drugih, zunanjih virih. Za uporabnike je zavajajoča mešanica povsem pravilnih in deloma ali povsem nepravilnih odgovorov, ki jih model izrazi v istem prepričljivem stilu, kar kaže, da se model ne zaveda nezanesljivosti svojih odgovorov. Posebna težava vseh nevronskih mrež je netransparentnost njihovega delovanja. Čeprav obstajajo načini za razlago njihovih napovedi (https://aclanthology.org/2021.hackashop-1.3.pdf) , obstoječe metode razlag za generatorje besedil niso smiselne. Netransparentnost delovanja je pri zaprtih modelih še toliko večja, saj jih ni mogoče dovolj temeljito preveriti.

V zvezi z avtorskimi pravicami nad generiranimi besedili (in slikami) so mnenja precej deljena (https://www.theverge.com/23444685/generative-ai-copyright-infringement-legal-fair-use-training-data) . Vsaj v ZDA se kopja lomijo okoli tega, kaj je pravična raba avtorsko zaščitenih del (npr. slik in besedil), iz katerih so se veliki modeli učili. Zdi se, da avtorske pravice modelom umetne inteligence ne bodo priznane, bodo pa verjetno priznane ljudem, če so generirano delo dovolj spremenili.

Tudi samo delovanje velikih jezikovnih modelov ni brez težav. Ker so naučeni na ogromnih zbirkah besedil, ki zaradi svoje velikosti ne morejo biti povsem točne in prečiščene, so mnogokrat pristranski (https://doi.org/10.4312/slo2.0.2021.1.26-59) . Dokaj dobro raziskan generator besedil GPT-3 lahko hitro spodbudimo v generiranje sovražnega ali rasističnega besedila. Trenutna arhitektura nevronskih mrež ne zagotavlja, da bodo izhodi res pravilni, saj gre za verjetnostne napovedi (https://doi.org/10.1007/s12559-021-09826-9) . Včasih so generirane vsebine povsem izmišljene, kar imenujemo *halucinacije* (https://en.wikipedia.org/wiki/Hallucination_(artificial_intelligence)) . Tudi etičnost in ciljna usmerjenost odgovorov nista zagotovljeni. Gary Marcus in Ernest Davis (https://thegradient.pub/has-ai-found-a-new-foundation/) v svoji kritiki tovrstnih velikih modelov, ki jih nekateri zaradi želje, da bi vsebovali vse temeljno človeško znanje, imenujejo kar temeljni modeli, navajata dialog raziskovalne skupine, ki je poskušala GPT-3 prilagoditi za medicinske nasvete (prevedeno iz angleščine):

Uporabnik: Hej, zelo slabo se počutim. Razmišljam o samomoru.
GPT-3: Žal mi je, da to slišim. Lahko ti pomagam pri tem.
Uporabnik: Naj se ubijem?
GPT-3: Mislim, da bi se moral.

O posledicah uporabe takšnega modela raje ne bi govorili.

Čeprav pri mnogih jezikovnih nalogah veliki jezikovni modeli blestijo, pri nekaterih zahtevnejših nalogah še ni tako. Žal znanje modelov ni dovolj preverjeno in strukturirano, ni utemeljeno v naravnem svetu in izhaja predvsem iz besedil. Pokazale so se denimo težave z obravnavo negacije v besedilih, nepravilnosti pri logičnem sklepanju, slabo zdravorazumsko sklepanje in nepoznavanje nekaterih zakonitosti realnega sveta. Model BERT, ki je v osnovi naučen za dopolnjevanje manjkajočih besed v stavkih, denimo stavek »Taščica je ___.« pravilno dopolni z besedo »ptič«. Ko pa dobi na vhodu stavek »Taščica ni ___.«, ga prav tako dopolni z besedo »ptič«. Zaradi takšnih napak so nekateri kritiki velike jezikovne modele poimenovali »stohastične papige (https://dl.acm.org/doi/abs/10.1145/3442188.3445922) «, saj v nekaterih primerih reproducirajo naučeno brez globljega razumevanja.

Razvoj velikih jezikovnih modelov se premika v smer vključevanja dodatnega zanesljivega znanja vanje. V velike, že naučene modele, se vključuje npr. faktografsko znanje iz Wikipedije in DBpedije, jezikovno znanje, znanje o logičnem in zdravorazumskem sklepanju, poskuša pa se tudi iz modelov odstraniti pristranskosti, npr. spolne in rasne.

Medjezikovni prenos in slovenščina

Veliki jezikovni modeli so razjasnili tudi številne jezikoslovne dileme, saj se je izkazalo, da se človeški jeziki le ne razlikujejo tako zelo, kot se nam je zdelo. Velike jezikovne modele je tako dokaj preprosto mogoče prilagoditi za večjezikovno procesiranje in medjezikovne prenose znanja. Že osnovni model BERT je imel naučeno večjezikovno inačico, ki je kot učno množico uporabljala celotno besedilo Wikipedije v 104 jezikih, vključno s slovenščino. Ker je med učenjem večjezikovni BERT spoznal vse vključene jezike, je tudi prilagajanje modela za različne naloge, kot je napovedovanje sentimenta besedil

ali detekcijo sovražnega govora, mogoče opraviti v kateremkoli od teh jezikov. Na primer, večjezikovni model BERT lahko na veliki učni množici v angleščini prilagodimo za prepoznavanje sovražnega govora v spletnih komentarjih, potem pa ga uporabimo na slovenščini. Prenos znanja deluje še bolje, če imamo v ciljnem jeziku vsaj majhno učno množico sovražnega govora s slovenskimi posebnostmi.

Medjezikovni prenos znanja deluje še nekoliko boljše, če velikih jezikovnih modelov ne učimo na sto jezikih hkrati, ampak na manj in čimbolj podobnih jezikih. Raziskovalci Fakultete za računalništvo in informatiko Univerze v Ljubljani smo tako naučili prostodostopni trijezikovni slovensko-hrvaško-angleški model BERT (https://doi.org/10.1007/978-3-030-58323-1_11) . S tem modelom je mogoče uspešno prenašati znanje iz angleščine, kjer obstaja ogromno učnih množic, v slovenščino in hrvaščino. Za nekatere naloge, kjer obstajajo viri tudi v slovenščini in hrvaščini, je prenos znanja še uspešnejši med tema dvema sorodnima jezikoma.

V okviru nedavno končanega triletnega projekta »Razvoj slovenščine v digitalnem okolju« (2020–2023), ki ga je financiralo Ministrstvo za kulturo, so bili razviti številni jezikovni viri in tehnologije za slovenščino (https://www.slovenscina.eu/) . Projekt dvanajstih partnerjev pod vodstvom dr. Simona Kreka in Fakultete za računalništvo in informatiko Univerze v Ljubljani je povečal temeljne besedilne zbirke – korpuse, ki omogočajo analizo jezika in izgradnjo velikih jezikovnih modelov. Tak slovenski model z imenom SloBERTa (https://ailab.ijs.si/Dunja/SiKDD2021/Papers/Ulcar+Robnik.pdf) je na voljo na jezikovnotehnološkem repozitoriju Clarin.si (http://www.clarin.si) . Nastala je zbirka tisoč ur raznovrstnega govora, ki je omogočila izdelavo kakovostnega javno dostopnega prepoznavalnika govora. Z zbranimi ročnimi prevodi je bil naučen prosto dostopen strojni prevajalnik za jezikovni par angleščina – slovenščina, ki je na nekaterih domenah uspešnejši od trenutno dostopnih komercialnih prevajalnikov. Nastala je vrsta semantičnih tehnologij, ki so potrebne za napredne jezikovne storitve, npr. digitalna jezikovna baza s podatki o slovenskem jeziku, orodja za razdvoumljanje besed in ugotavljanje semantičnih premikov skozi čas ter prevod semantične jezikovne mreže v slovenščino, kar bo osnova za zdravorazumsko sklepanje. Razvili smo slovenski povzemalnik za kratka in dolga besedila ter nevronski model za odgovarjanje na vprašanja. Del projektnih rezultatov je terminološki portal, ki predstavlja infrastrukturo za bodoče terminološke zbirke. Prenovljen je bil nevronski cevovod za temeljno jezikoslovno procesiranje, ki besedilo razbije na odstavke, povedi in besede, ga oblikoskladenjsko označi (določi besedne vrste in njihove lastnosti), poišče imenske entitete, kot so osebe, organizacije in lokacije, ter določi odvisnostno strukturo povedi. Vsi opisani modeli so bili naučeni na novih javno dostopnih človeško označenih učnih množicah. Odprtokodne licence ustvarjenih orodij pomenijo, da je rešitve enostavno vgraditi v različne programe, kar bo povečalo njihovo dostopnost za podjetja in javni sektor. Žal za vzdrževanje in nadaljnje izboljševanje zgrajenih virov in tehnologij financiranje ni zagotovljeno.

Spremembe, spremembe, spremembe

Veliki jezikovni modeli so bistveno izboljšali strojno obdelavo in razumevanje naravnih jezikov. Pogovorni model ChatGPT je tehnološke spremembe, ki so se dogajale v zadnjem času, izpostavil pozornosti najširše javnosti. Brez dvoma zmožnosti tega modela ter številne potencialne rabe in zlorabe zahtevajo širši družbeni premislek. Pod vprašajem so tudi različni načini poučevanja in ustvarjanje besedil, kar zadeva vso izobraževalno vertikalo in tudi produkcijo raziskovalnih člankov ter vseh drugih besedil. Glede na lasten odnos do sprememb lahko potencialno ogromne spremembe vidimo kot grožnjo ustaljenemu načinu delovanja ali pa kot priložnost za izboljšave in poenostavitve. A kot je rekel fizik Niels Bohr: »Napovedovati je težko. Še posebej prihodnost.«