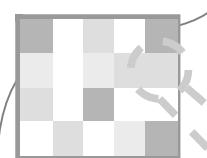


Recept na model, který pohání ChatGPT

1) Vezmi *Transformer* a nastav mu náhodně parametry. (Pokud chceš hodně schopný model, vezmi opravdu velký *Transformer*.)

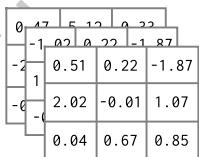
Nenatrénovaný jazykový model



Transformer

Základ dnešních jazykových modelů, jako je např. ChatGPT. Několik miliard čísel (tzv. parametrů) a sada vzorců, které určují, jak parametry smíchat se vstupem a získat tím výstup – třeba příští slovo).

2) Nasbírej spousty textů: z internetu, z dokumentů, z knih. Čím více textů a čím kvalitnějších, tím lépe.



Všechny znalosti a schopnosti modelu jsou uloženy v obrovských sadách parametrů naučených z dat.

Backpropagation

Algoritmus, který umí zjistit, které parametry modelu způsobily předpověď modelu a jakou mírou se na předpovědi podílely. Díky tomu můžeme následně zvýšit nebo snížit hodnotu parametrů odpovídajícím způsobem a tím model učit.

3) Pro každý kus textu se zeptej modelu, které slovo by mělo následovat. Když se netrefí do správného slova, lehce uprav hodnoty příslušných parametrů tak, aby slovo pro tento kus textu předpovídalo s vyšší pravděpodobností.

Dům stál na mírném svahu na samém konci vesnice.
Stal o samotě, s výhledkou na šíré lány (...)



4) Hotovo! Náš model umí obstojně předvídat další slovo ve větě v textech, na kterých byl natrénovaný.

po několika kvadrilionech výpočetních operací



Natrénovaný jazykový model

5) Bonus: pokud chceš, aby uměl jazykový model fungovat jako schopný asistent...

a) Dotrénуй ho na textech, ve kterých se plní instrukce.

uživatel: Máš to špatně!
odpověď #1: No to teda nemám!
odpověď #2: Moje chyba, omolouvám se, ...

b) Dotrénuj ho, aby preferoval odpovědi, které se lidem líbí.

Generování příštího slova ve větě

Embedding

0.8 -0.1 1.6 0.2 -0.5

Sada několika set až tisíc čísel, který reprezentuje token (=slovo nebo skupinu znaků) ze vstupu.

Feed-forward

Vrstva, která umožňuje modelu zpracovávat reprezentaci jednoho konkrétního tokenu.

Výpočetní operace

Každá šipka symbolizuje operace nad reprezentacemi tokenů. Uvnitř Transformer bloků jde především o násobení a sčítání, při kterém se kombinují vstupní data a parametry modelu.

Hlavní
14258

město
1572

Francie
522

je
1014

Jazykový model

attention

feed-forward

Transformer blok

Vrstvy attention a feed-forward tvoří tzv. Transformer blok, který se v typickém modelu opakuje cca 8-64x. Bloky postupně operují na reprezentaci vstupu a přetvázejí ji v reprezentaci výstupu.

opakuj
Nx

A můžeme jít generovat text!

Token id

1013: si
1014: je
1015: \$\$

Slovník typického modelu obsahuje několik desítek tisíc tokenů. Každý token má svoje unikátní "idčko".

Attention

Vrstva, která přimíchává do reprezentace tokenu informaci z ostatních tokenů.

Dekódovací algoritmus

Rozhoduje, který token vygenerovat na základě pravděpodobnosti přiřazených jednotlivým tokenům ve slovníku.

Co myslíš, chceme vždy token s nejvyšší pravděpodobností?

Paříž

a ...
... Paříž ... zyota
pravděpodobnosti všech tokenů, které má model ve slovníku

Paříž

Generování textu: otázky a odpovědi

Umí si jazykové modely i dohledávat odpovědi na internetu?



Krátká odpověď: Neumí, ale někdo to může dělat za ně.

Dlouhá odpověď: Jazykový model jako takový hledat na internetu neumí. Všechno, co umí, se musel naučit během trénování.

V dnešní době ale mají komerční jazykové modely typicky externí vyhledávací modul. Tento modul hledá informace související s dotazem a předkládá je jazykovému modelu na vstupu společně se vstupem od uživatele. Model pak s nimi pracuje stejně, jako se zbytkem textu. Kvalita těchto výsledků může ovlivnit vygenerovaný text jak pozitivně, tak negativně.

Proč modely neumí ostatní jazyky tak dobře, jako angličtinu?



Krátká odpověď: Protože angličtina je jazykem internetu.

Dlouhá odpověď: Zatím nedokážeme učit modely jazyk ani zdaleka tak efektivně, jako třeba děti. Ani celá Wikipedie nestáčí na to, aby se velký jazykový model naučil modelovat jazyk bez chyb.

Schopnosti modelu proto rostou s množstvím textů, na kterých byl trénovaný. A textů v angličtině se dá najít opravdu hodně! Překvapivě i v češtině jich je celkem dost, ale třeba v irštině nebo telugu bude model pokulhávat.

Co dělá velké jazykové modely tak inteligentní?



Krátká odpověď: To je mi ale těžká otázka...!

Dlouhá odpověď: Modely do nějaké míry umějí zopakovat to, co viděly v trénovacích datech. Chytrá odpověď proto mohla jen "ležet na internetu".

Modely ale umí částečně i generalizovat: kombinovat naučené vzory v originální odpovědi. Důvodem může být to, že s velkým množstvím dat může být generalizovat jednodušší, než se učit vše nazepaměť.

A tady už přicházejí další otázky: Co dělá inteligentní člověka? A co je to vlastně ta "intelligence"?

Má model seznam všech slov ve všech jazyčích?



Krátká odpověď: Ne – ale jejich částí ano!

Dlouhá odpověď: Všech slov by bylo opravdu až příliš (už jen všechny vyskočované tvary slov v češtině!). Proto používáme jako tokeny tzv. "subwordy": slova a části slov, ze kterých můžeme libovolná slova poskládat. Frekventovanější slova máme ve slovníku přímo, ta méně častá složíme z více částí.

Rozsekat text na subwordy a pak ho zase poskládat je úkolem specializovaného algoritmu, tzv. tokenizéra, který pracuje nezávisle na jazykovém modelu.

Jak energeticky náročné je natrénovat model? A kolik energie spotřebuje jeden dotaz?



Krátká odpověď: Trénovat je náročné, generovat ne tolik.

Dlouhá odpověď: Záleží na velikosti modelu, ale odhaduje se, že natrénovat model s 175 mld. parametry stojí 1.2 GWh energie, což odpovídá roční spotřebě 120 amerických domácností. Modely se ale naštěstí netrénují tak často.

Generovat z natrénovaného modelu text už je mnohem efektivnější: průměrná odpověď ChatGPT spotřebuje cca 0,3 Wh, což odpovídá např. rychlovárné konviči zapnuté na 1 sekundu.

[1] <https://blog.samaltman.com/the-gentle-singularity>

[2] <https://adasci.org/how-much-energy-do-langs-consume-unveiling-the-power-behind-ai/>

[3] <https://andymasley.substack.com/p/a-cheat-sheet-for-conversations-about>

Kde se můžu naučit o jazykových modelech víc?



Krátká odpověď: U nás na ÚFALu!



Dlouhá odpověď: Podobnými tématy se zabýváme na Ústavu Formální a Aplikované Lingvistiky (ÚFAL), což je součást Matematicko-fyzikální fakulty Univerzity Karlovy.



Na naší katedře propojujeme znalosti o jazyce se znalostmi ze strojového učení. Učíme o tom bakalářské i magisterské předměty, jako třeba Deep Learning nebo Large Language Models.