⚡️🚀המאמר היומי של מייק 04.07.24:⚡️🚀

How Do Large Language Models Acquire Factual Knowledge During Pretraining?

המאמר חוקר נושא מתי מודלי שפה אשכרה רוכשים ידע עובדתי (למשל שעיר בירה של צרפת היא פריס) במהלך אימון מקדים. בנוסף המאמר גם בודק כמה זמן לוקח לשכוח ידע עובדתי. אוקיי, אתם בטח זוכרים שאנו מאמנים מודלי שפה שלנו עם אחת הצורות של משפחת מורד הגרדיאנט (gradient descent או GD). בד״כ דוגמים כמה דוגמאות הסט האימון שלנו (מיני-באץ') ומזיזים לינארית את משקלי המודל לכיוון הנגדי של הגרדיאנט הממוצע של מיני-באץ'.

המאמר בונה דוגמא של טקסט המכיל ידע עובדתי ומכניס אותו למיני-באץ' כל כמה איטרציות של GD. המחברים מצאו כמה דברים מעניינים. למשל כמות דאטה שהמודל אומן עליו לפני התחלת הזרקת ידע עובדתי לא משפיע על מספר האיטרציות הנדרש ללמידה של ידע עובדתי. כלומר יותר ״ידע״ הנמצא כבר במודל לתורם למהירות הלמידה.

שנית, המאמר מראה שמהירות הלמידה של ידע עובדתי לא מושפעת ממתי מתחילים להזריק למודל את הידע. כלומר מודל מאומן לאו דווקא תלמיד יותר טוב. ויש עוד כמה תגליות מעניינות במאמר.

איך בודקים האם המודל אכן למד את הידע העובדתי שהזרקנו - המחברים לא מרחיבים על כך אבל כנראה זה מחושב דרך likelihood של התשובה הנכונה על השאלה לגבי פיסת ידע עובדתי זה, למשל ״מה עיר הבירה של צרפת״.

https://arxiv.org/abs/2406.11813