⚡️🚀המאמר היומי של מייק -31.10.24: ⚡️🚀  
Understanding Transformers via N-gram Statistics

מאמר די נחמד ולא רגיל מבית גוגל. המאמר מחזיר אותנו לתקופה שלא מידלנו את השפה הטבעית באמצעות מודלים סטטיסטיים עם עשרות ומאות מיליארדי פרמטרים. פעם ניסינו להשתמש ב- n-grams כדי לשערך את ההתפלגות של המילים בטקסט. כמובן גישות כאלו לא יכולות לעבוד עבור דאטהסטים בעל עשרות טריליוני טוקנים כמו שיש לנו היום אבל אולי אפשר לקחת LLMs גדולים ולבדוק האם ניתן לקרב את חיזויהם באמצעות סטטיסטיקות על n-grams. כדי לא לסבך המאמר לא בודק את זה על למידת in-context.

וזה בדיוק מה שהמאמר הזה (שיש לו רק מחבר אחד שזה די נדיר בימינו) עושה. הוא בודק האם ניתן לחזות את הטוקן הבא שמודל שפה מאומן חוזר באמצעות סטטיסטיקה של n-grams שבאים לפניו בטקסט. במקרה הזה n-grams בנויים לא ממילים אלא מטוקנים. דרך אגב הסטטיסטיקה של n-grams אינה חייבית לכלול את כל n הטוקנים הבאים לפני הטוקן הנחזה אלא עשויה ״להכיל חורים״(כלומר יכולה לקחת טוקן i-1, i-2 i ו- i-4 עבור 3-gram - נצטרך למצע מעל טוקן i-3 בשביל כך).

המחבר מצא כמה דברים מעניינים. ניתן לשערך את החיזוי של מודל שפה עם 7-gram (עבור דאטהסטים שהם בחרו) בלא מעט מקרים. בנסוף נמצא כי לטוקנים בעל שונות נמוכה (של ההתפלגות שלהם) n-grams מצליחים יותר מאשר לטוקנים בעל שונות חיזוי גבוהה. מעניין שככל שמאמנים מודל שפה יותר יותר קשה לקרב אותה עם n-grams (צריך להגדיל את n או לא משנה מה ה-n דיוק הקירוב יורד).

אהבתי…

https://www.arxiv.org/abs/2407.12034