המאמר היומי של מייק - 11.04.25  
 Arithmetic Without Algorithms: Language Models Solve Math with a Bag of Heuristics

כבר סקרתי בעבר כמה מאמרים על מודלי שפה לחישוב נוסחאות אריתמטיות המכילות פעולות חשבוניות סטנדרטיות כמו פלוס, כפול וכדומה. לדעתי מודלי שפה פחות מיועדים למשימות מהסוג הזה (יש לנו מחשבונים, בפייטון וכאלו) אבל בכל זאת יש מחקרים מעניינים בנושא הזה. ויש סיבה נוספת לבחירת המאמר הזה - הוא נכתב על ידי חוקרים ישראלים ותמיד נהנה לסקור תוצרת מקומית.

אז כאמור המאמר חוקר מה קורה בתוך מודל הטרנספורמר כאשר מודל שפה מקבל משימה אריתמטית. למעשה המחברים מנסים לאתר מה שנקרא נתיב החישובי (circuit) בתוך הטרנספורמר כלומר רכיביו המבצעים בפועל את "החישובים הנדרשים" עבור משימה זו. אתם בטח זוכרים שבלוק טרנספורמר מורכב משתי שכבות עיקריות (יש גם שכבות נרמול) שהם מנגנון attention מרובה ראשים או MHA ושכבת MLP המורכבת משתי שכבות לינאריות ואקטיבציה לא לינארית ביניהם. אז הנתיב החישובי מורכב מנוירונים מסוימים בתוך ה-MHA או בתוך ה-MLP.

כדי לאתר את הנתיב החישובי, המחברים מבצעים החלפת אקטיבציות (activation patching) של נוירונים בתוך הטרנספורמר המאפשרים לשערך את החשיבות של שכבות MLP וכל ראשי attention בכל מיקום בסדרת קלט (פרומפט אריתמטי). איך עושים זאת? לוקחים פרומפט אריתמטי מסוים (לדוגמא, "226 − 68 ="), ופרומפט אקראי שמוביל לתוצאה שונה (למשל, "21 + 17 ="). לאחר חישוב של אקטיבציות המודל עבור הפקודה האקראי, מזינים את פרומפט המקורי למודל.

בשלב זה מתערבים בחישוב (patching) — כלומר, מחליפים את אקטיבציות של שכבת MLP בודדת או ראש attention באקטיבציות שחושבה מראש עבור הפרומפט האקראי. בהמשך בודקים כיצד ההתערבות משפיעה על ההסתברויות של שני הטוקנים של התשובות(עבור הפרופמט המקורי ועבור האקראי) - יש נוסחה שמשערכת השינויים בטוקני התשובות. לאחר מציאת הנתיב החישובי עבור הדוגמאות השונות המאמר משערך את ״נקיונם״ על ידי החלפה של כל האקטיבציות באקטיבציות ממוצעות על פני דאטהסט גדול של פרומפטים אריטמתיים כאשר רק האקטיבציות של הנתיב החישובי נותרו על כנם. המחברים הראו שהחלפה זו כמעט ולא משפיע על הלוגיטים של התשובה הנכונה.

אחרי מציאת נתיבים חישוביים אלו המחברים ניסו להבין איזה משמעות אריתמטית יש להם. כתוצאה מכך התבררה תמונה די מעניינת. המחברים הראו כי הפעולות של נתיבים אלו הם למעשה יוריסטיקות שונות המאפשרות לפתור את התרגיל. למשל היו נוירונים שמטרתם היא להגיד האם התוצאה נמצאת בתחום [150, 180] או שהתוצאה מתחלקת ב-5. שילוב של שערוכים אלו מאפשר למודל לפתור תרגילים אריתמטיים פשוטים יחסית הלא מערבים מספרים גבוהים מדי. זה די מסביר למה LLMs מתקשים עם פעולות על מספרים גבוהים.

בנוסף יש כמה מציאות מעניינות. רוב החלקים הבולטים של הנתיבים החישוביים נמצאים בשכבות MLP ולא בראש attention. הדבר המעניין השני הוא העובדה שהמודל ״די מתכנס״ לתשובה הנכונה כבר בשכבות הביניים (ניתן להפיק אותה משם על ידי שכבה לינארית).

https://arxiv.org/abs/2410.21272