המאמר היומי של מייק: 14.05.25  
ICLR: In-Context Learning of Representations

מודלי שפה מסוגלים לעשות הרבה יותר מאשר רק לשחזר עובדות או לבצע הוראות אלא הם מסוגלים להתאים את הייצוגים הפנימיים שלהם בהתבסס על ההקשר בלבד ללא עדכון משקליו(למידה in-context). המאמר שנסקור היום מראה כי מודלי שפה יכולים לארגן מחדש את הגיאומטריה הסמנטית הפנימית שלהם באופן מלא, רק באמצעות פרומפט כאשר משקליו נותרים ללא שינוי. וזה לא שינוי "שטחי" בפלט. מדובר בארגון מחדש של מרחב הייצוג הפנימי של המודל, שנוצר מתוך מבנה ההקשר בלבד.

אימון מקדים של מודל שפה בונה מרחבים סמנטיים יציבים: מילים נרדפות מתקרבות זו לזו, מדינות יוצרות קבוצות גיאופוליטיות, וימי השבוע נפרשים במעגל. אבל מה קורה כאשר פרומפט משנה את היחסים האלו? האם המודל יכול לבנות משמעות חדשה רק מההקשר? זו בדיוק השאלה שבודקים במאמר. הם מסירים את הרמזים הסמנטיים שקיימים במודל מהאימון המקדים וגורמים למודל להסיק את המשמעות אך ורק מתוך מבנה הסדרה בפרומפט, ומגלים בכך יכולת מפתיעה של המודל ללמוד גיאומטריה ייצוגית חדשה בתוך ההרצה.

במערך הניסוי, בונים גרף שכל קודקוד בו הוא טוקן מוכר (מילה כמו תפוח או רכבת). הליכה אקראית על הגרף מייצרת סדרות של טוקנים שהוא הפרומפט. המודל מתבקש לחזות את הצעד (קודקוד) הבא, למרות שלמילים עצמן אין רמזים סמנטיים ישירים. למשל תפוז יכול להיות צמוד לרכבת בגרף ומאוד רחוק (מספר הקשתות המינימלי ביניהם) לאגס.

כדי להצליח בחיזוי קודקוד הבא, המודל חייב לחשוף את מבנה הגרף ולהתאים מחדש את הייצוגים של הטוקנים בהתאם. המבנה חבוי בסדרות מילים הנדגמות מהגרף ולא במילים עצמן. כאשר מנתחים את האקטיבציות הפנימיות בשכבות הטרנספורמר, מגלים תופעה מרתקת. בתחילה (עבור סדרות קצרות הנדגמות מהגרף המוזנות למודל כפרומפט), הייצוגים של הטוקנים עדיין משקפים את המשמעות מהאימון המקדים. אך ככל שההקשר מתארך (דגימות ארוכות יותר), המרחב משתנה בפתאומיות: טוקנים סמוכים בגרף מתקרבים זה לזה במרחב הייצוגי.

זה לא תהליך הדרגתי אלא קפיצה חדה, סוג של שינוי פאזה. מתחת לאורך הקשר קריטי (מספר דוגמאות מגרף המילים שלנו), המשמעות המקורית שולטת. ברגע שעוברים את הסף, המודל "מתמסר" למבנה החדש, וממפה את העולם הפנימי שלו לפי המבנה החבוי בפרומפט. התופעה הזו מראה שהמודל לא פשוט משנן זוגות טוקנים. הוא לא רק משחזר מילה הבאה מה"זכרון" אלא בונה מבנה עקבי ומקיף (החל מהשכבות הדי מוקדמות של הטרנספורמר) מתוך דפוסים מקומיים. אחת טענות המאמר אומרת כי "שינויים שטחיים" שמתבססים על שינון בלבד לא מצליחים להשתוות לביצועים של המודל או לשחזר את הגיאומטריה שנוצרת.

ארחיב על כך. נניח שבנינו גרף שבו קודקודים הם מילים ומשקלי הקשתות הם מרחקים בין ייצוגי המילים על יד המודל (נניח על ידי שכבה מסוימת). אז מתברר שהגרף הזה איזומורפי (ספקטרלית) לגרף של מילים שאנו דוגמים ממנו לפרומפט. כלומר אם נחשב את ה-pc (שזה principal component או הכיוון המוביל או מתאים לערך העצמי הגבוה ביותר) של הייצוגים הספקטרליים (שזה בגדול מטריצה שכנויות ממושקלת) של שני הגרפים נקבל גרפים דומים.

כלומר אם נבנה עבור כל אחד מהם גרף שבו המרחק (שזה 1 חלקי המשקל) בין שני קודקודים(מילים) מוגדר על ידי המרחק בין המימד הראשון בוקטור pc המתאים לקודקודים אלו, נקבל גרפים דומים. כלומר מטריצות שכנויות של של שני הגרפים (הראשון מייצג מרחק בין ייצוגי המילים על ידי המודל השני גרף השכניות שממנו דוגמים לפרומפט) הם די דומים שזה די מדהים. כלומר יצוגי המודל אשכרה לומדים את ״עיקרי גרף השכניות בפרומפט״

המחברים גילו עוד משהו מעניין. כאשר משתמשים במילים בעלי משמעות סמנטית חזקה (כמו ימות השבוע), המודל לא מוחק אותם. במקום זאת, הוא שומר את מבנה השכניות הקודם במימדי pc הראשונים (בגרף הבנוי על ייצוג המילים על ידי המודל), ומטמיע את המבנה החדש במימדים הבאים (אבל משמעותיים) של המרחב הייצוגי. כך, המודל מצליח להחזיק בו זמנית משמעות מוקדמת ומשמעות חדשה, על ידי הפרדה גאומטרית בתת-מרחבים שונים.

המחברים משווים את המעבר הזה לתופעת פרקולציה בפיזיקה: חיבורים מקומיים מצטברים עד שנחצה סף קריטי, ואז מופיעה פתאום תבנית כוללת. כאן, אורך ההקשר, לא גודל המודל, הוא שמכתיב את הופעת המבנה החדש. ככל שהפרומפט מתארך, כך המבנה הפנימי נעשה צפוי יותר, עד שמתחוללת קפיצה פתאומית במבנה. המחקר הזה משנה את ההבנה שלנו לגבי למידה in-context. הוא מראה שמודלים לא רק משנים תגובות להקשר אלא בונים מחדש את עולמם הפנימי לפי דרישות ההקשר.

https://arxiv.org/abs/2501.00070