**מצילומי רנטגן סטטיים לסריקות MRI חיות: שיטה חדשה לצפות באיך LLMs לומדים, סקירה 517**

**מאמר היומי של מייק: 30.09.25**

**EVOLUTION OF CONCEPTS IN LANGUAGE MODEL PRE-TRAINING**

במשך שנים, הבנת האופן LLMs לומדים הרגישה כמו חקר האסטרונומיה לפני המצאת הטלסקופ. יכולנו לראות את התוצאות המפוארות, קבוצות הכוכבים המבריקות של יכולות מתהוות, והיו לנו תיאוריות מתמטיות מבוססות שתיארו את מבנה היקום בגדול. עם זאת, התהליך הבסיסי ההיווצרות של כוכב, או במקרה זה, של מושג, נותר אפוף ב"קופסה שחורה" מתמטית בלתי מובנת. היה לנו את ה"לפני" (אתחול אקראי) ואת ה"אחרי" (מודל מאומן), אבל ה"במהלך" המכריע נותר תעלומה.

מאמר שנסקור היום מציג טלסקופ חדש ועוצמתי. הוא מציג מתודולוגיה חדשנית אשר, לראשונה, מספקת מבט רציף ברמת התכונה (feature) אל תוך דינמיקת הלמידה של LLM. הוא מסיט את הבנתנו מסדרה של תמונות סטטיות לנרטיב זורם ומתפתח של האופן שבו מושגים מופשטים צצים מתוך סטטיסטיקה גולמית. תרומתו המרכזית של המאמר אינה ממצא ספציפי, אלא יצירת עדשה מסוג חדש להתבוננות בתהליך עצמו.

הקושי המרכזי בחקר אימון LLMs היה בעיית ההמשכיות. חוקרים בתחום הפירוש המכניסטי (mechanistic interpretability) הפכו מיומנים בניתוח מודלים מאומנים באמצעות כלים כמו אוטו-אנקודרים דלילים (SAEs). שיטות אלו פועלות כמו מנסרה, ומפרקות את האקטיבציות הפנימיות הצפופות של המודל ל"מילון" של פיצ'רים נקיים וניתנים לפירוש כלומר מושגים כמו "שמות עצם ברבים", "ז'רגון משפטי" או "תחביר JSON". זה דומה לביצוע נתיחה שלאחר המוות: זה אומר לך הכל על המצב הסופי, אבל שום דבר על האופן שבו הוא נוצר.

אי אפשר פשוט להריץ SAE על מודל בצעד 1,000 של האימון ושוב בצעד 10,000 ולהשוות ביניהם. מדוע? כי אין ערובה לכך ש"תכונה #47045" מהניתוח הראשון מתאימה בכל דרך שהיא ל"תכונה #47045" מהניתוח השני. המודל מארגן את עצמו מחדש ללא הרף; מושגים עשויים לשנות את הייצוג שלהם, או "להסתובב" במרחב האקטיבציה הרב-ממדי. ללא מסגרת התייחסות יציבה, אתה לא עוקב אחר אבולוציה; אתה פשוט מסתכל על סדרה של תמונות לא קשורות.

### החידוש: למידת מילון חוצת-תמונות-מצב (Cross-Snapshot)

התובנה המכריעה של המחברים היא להתאים טכניקה הנקראת crosscoders. במקור, הם תוכננו כדי ליישר ולהשוות תכונות בין *שכבות* שונות של מודל יחיד, אך החוקרים הסבו את השימוש בגישה זו כדי לנתח אקטיבציות מצ'קפונים שונים של האימון. במקום לאמן מילון נפרד עבור כל snapshot(מצב המודל אחרי צ'קפוינט) הקרוסקודר מאומן לפרק בו-זמנית את האקטיבציות מ*כל* סנפשוט שנבחרו באמצעות מרחב פיצרים(המופקים באמצעות האנקודר של הקרוסקודר) יחיד ומאוחד. ואז מעבירים את הפיצרים האלו בחזרה באמצעות הדקודרים השונים לכל סנפשוט. ד״א המאמר מכיל טעות לכאורה בהגדרה של האנקודר במקום אחד ועשה אותו תלוי בסנפשוט (צעד האימון).

דמיינו את המקודד כמתרגם אוניברסלי של מושגים. הוא מסתכל על המצב הפנימי של המודל, בין אם בצעד 100 או 100,000, ומפעיל פיצ'ר ספציפי ועקבי ממילון הפיצ'רים המאוחד שלו, ואומר, "המושג 'x' נוכח כאן". הקסם קורה במפענחים. כל סנפשוט מקבל מפענח משלה, שמשחזר את האקטיבציה המקורית של המודל באמצעות התכונה שזוהתה על ידי המקודד האוניברסלי. בסנפשוט מוקדם, לפני שהמודל למד על פיצ'ר x, המפענח הספציפי שלו עבור פיצ'ר x יהיה חלש (כלומר המקדם שלו בדקורר). ערך המילון שלו עבור אותו מושג הוא למעשה ריק. ככל שהמודל מתאמן והמושג הזה מתגבש, המפענח של אותו סנפשוט מאוחר יותר מפתח ייצוג חזק ומדויק עבורו.

תכנון זה מספק את זווית מתודולוגית הבאה: עוצמת הווקטור של כל פיצ'ר במפענח הספציפי לכל סנפשוט משמשת כפרוקסי ישיר וכימותי לנוכחות ולבשלות של אותה תכונה באותה נקודת זמן. זה יוצר את מסגרת ההתייחסות היציבה שהייתה חסרה. זה מאפשר לחוקרים לשרטט את "סיפור החיים" של כל פיצ'ר ופיצ'ר, ולצפות בהם צצים, מגיעות לשיא ודועך לאורך האימון.

### התוצאות: מחזה בשני שלבים של למידה

באמצעות העדשה החדשה הזו, החוקרים הבחינן בתהליך למידה מובהק בן שני שלבים. בשלב הראשון, אותו הם מכנים "שלב הלמידה הסטטיסטית", הירידה המהירה בלוס של המודל מונעת כמעט אך ורק מלמידת דפוסים סטטיסטיים גסים, כמו התפלגויות של יוניגרמות וביגרמות. בשלב זה, הייצוגים הפנימיים של המודל צפופים, ומעט פיצ'רים מובחנים וניתנים לפירוש נוצרין. לאחר מכן, בנקודת מפנה עקבית להפליא סביב צעד 1,000, מתחילה המערכה השנייה: "שלב למידת התכונות". כאן, הגיאומטריה הפנימית של המודל עוברת שינוי דרסטי, ואלפי מושגים משמעותיים מתחילים להיווצר. המאמר מראה היררכיה ברורה בהופעה זו: פיצ'רים פשוטים יותר (כמו אלו המזהות את הטוקן הקודם ברצף) נוצרים מוקדם, בעוד שפיצ'רים מופשטים ומורכבים יותר (כמו אלו הנדרשות לאינדוקציה ולמשימות תלויות-הקשר) מופיעים בשלב מאוחר הרבה יותר באימון. המחברים מקשרים באופן סיבתי את הולדתן של פיצ'רים אלו ליכולות של המודל במשימות downstream, ומראים כיצד ביצועים במשימות כמו התאמת נושא-נשוא מתממשים רק לאחר שהתכונות הספציפיות וההכרחיות למיומנות זו נוצרו.

### סיכום

שיטה זו מספקת גשר חיוני בין שני סולמות מחקר נפרדים ב-AI. ברמה המיקרוסקופית, יש לנו חוקרי מעגלים (circuits) שממפים את המסלולים החישוביים המדויקים במודלים סטטיים. ברמה המקרוסקופית, יש לנו תיאורטיקנים שמתארים למידה באמצעות מעברי פאזה בסקייל גבוה. מאמר זה מספק את החוליה המזוסקופית החסרה. הוא מציע תיאור ישיר ואמפירי של האופן שבו מתרחש מעבר הפאזה המקרוסקופי: ניתן ממש לצפות כיצד המודל שולט תחילה בדפוסים סטטיסטיים גסים, ולאחר מכן, בנקודת מפנה מובחנת, מתחיל לאכלס את המילון המאוחד שלו באלפי תכונות חדות ומשמעותיות.

https://www.arxiv.org/abs/2509.17196