המאמר היומי של מייק - 03.03.25  
 The Geometry of Concepts: Sparse Autoencoder Feature Structure

המאמר חוקר את האופן שבו מקודדים אוטומטיים דלילים(SAE) מייצגים ומבנים מושגים ב-LLMs. החוקרים מנתחים את המבנה הזה בשלושה קני מידה היררכיים: אטומי, מוחי וגלקטי. המחקר מנסה לעשות לא מעט הקבלות בין מרחב האמבדינג של מודלי שפה למבנה המוח אבל כמובן זה לא אוצר ש-LLMs חושבים בדומה לנו.

מתודולוגיה:

נרענן כי SAEs הינו כלי לחקר של interpretability של LLMs. הם מאומנים לשחזר אקטיבציות של שכבה ספציפית במודל תוך שימוש בתת-קבוצה קטנה של פיצ'ירים שלהן בלבד. אילוץ דלילות זה מכריח את SAE לייצג כל נוירון בתור צירוף לינארי של מספר קטן של פיצ'רים סמנטיים שכל אחד מהם(פיצ'רים) מקודד מושג מסוים(ניתן לפרשנות). כלומר ה-SAE לומד מילון של וקטורי פיצ'רים(אמבדינגס) שבו כל נוירון מופעל באופן סלקטיבי עבור דפוסים סמנטיים מסוימים.

החוקרים משתמשים ב-SAEs כדי לחלץ פיצ'רים סמנטיים מייצוגים של מושגים ב LLMs. המחקר מתמקד בניתוח המבנה הגיאומטרי של ייצוגים אלה בשלושה קני מידה.

כדי לחשוף את המבנה הזה, החוקרים משתמשים ב-LDA כדי להסיר כיווני ״הסחה״ גלובליים במרחב האמבדינג, כמו אורך מילה, שעלולים לטשטש קשרים מושגיים סמנטיים. שלב זה חיוני במיוחד עבור הרמה האטומית, שבה יחסים אנלוגיים הופכים ברורים יותר לאחר הסרת ההשפעות המסיחות.

רמה אטומית: "גבישים" ותבניות גיאומטריות

בקנה המידה הקטן ביותר, המחקר מזהה "גבישים"- מבנים גיאומטריים כמו מקביליות טרפזואידים - בתוך מרחב התכונות הרב-ממדי. מבנים אלו מכלילים יחסיים ידועים כמו (גבר - אישה) :: (מלך - מלכה). החוקרים מציינים כי איכות הדפוסים הגיאומטריים משתפרת משמעותית כאשר מסירים כיווני הסחה גלובליים, כמו אורך מילים, באמצעות (LDA – Linear Discriminant Analysis).

### רמה מוחית: מודולריות מרחבית ו"אונות"

בקנה מידה בינוני, המחקר חושף מודולריות מרחבית בתוך מרחב פיצ'רים של ה-SAE. פיצ'רים השייכים לתחומים ספציפיים, כמו מתמטיקה וקוד, מקובצות יחד ליצירת "אונות" נפרדות, בדומה איזורים תפקודיים הנצפים ב-fMRI של המוח האנושי. החוקרים משתמשים במדדים שונים כדי לכמת את הלוקליות המרחבית של האונות ומגלים כי פיצ'רים מופיעים יחד בצפיפות גבוהה יותר ממה שהיה מצופה בגיאומטריית של פיצ'רים אקראיים . ממצאים אלה מצביעים על כך שה-SAE מארגן פיצ'רים קונספטואליים באופן המשקף התמחות תפקודית.

### רמה גלקטית: מבנים בקנה מידה רחב והתפלגות ערכים עצמיים

בסקאלה הגדולה ביותר, המחקר מגלה כי פיזור ענן הנקודות המאפיינות הוא אניזוטרופי(שונה בכיוונים שונים) מאופיין על Power Law של ערכים עצמיים, השינויים התלולים (בין ע״ע) ביותר שנצפים בשכבות האמצעיות של הרשת. הדבר מצביע על כך שהמורכבות והווריאציה של ייצוגי דאטה אינם אחידים בין השכבות, כאשר השכבות האמצעיות קולטות וריאציות עדינות יותר בנתונים. המחברים גם מנתחים כיצד משתנה האנטרופיה של אשכולות(בענני נקודות) בין השכבות השונות של המודל, ומספקים תובנות על המבנה ההיררכי של ייצוג מושגים בתוך המודל.

https://arxiv.org/abs/2410.19750