המאמר היומי של מייק: 26.06.25  
Open Problems in Mechanistic Interpretability

אינטרפרטביליות מכניסטית היא אולי תחום השאפתני ביותר כיום להבנת איך בינה מלאכותית באמת עובדת. לא מדובר כאן בהסברים בנפנופי ידיים או בהדגשות צבעוניות של חלקי טקסט אלא בהנדסה לאחור(reverse engineering) של הרשתות עצמן. הבנה אמיתית של איך רשת נוירונים פותרת בעיה: מהם החלקים הפנימיים שפועלים, באיזה סדר, באיזו לוגיקה, ואיך בדיוק הם מייצרים הכללה.

הגישה המרכזית שמוצגת במאמר מבוססת על שלושה שלבים: פירוק הרשת לרכיבים קטנים (בין אם אלו נוירונים, תתי-מרחבים או מעגלים), תיאור התפקיד הפונקציונלי של כל אחד מהם, ואימות כלומר בדיקה האם ההסבר שלנו באמת חוזה התנהגות, ואם כן עד כמה. כל אחד מהשלבים האלה מתגלה כקשה הרבה יותר ממה שנדמה.

הבעיה הבסיסית היא שפירוק לפי מבנה הארכיטקטורה של הרשת כלומר שכבות, נוירונים, ראשי attention פשוט לא עובד. החלקים האלה לא מתאימים למה שהרשת באמת מחשבת. נוירונים הם פוליסמנטיים(רב משמעים), תפקידים מתפרשים על פני שכבות שונות, ותכונות לא שוכנות בוקטור בודד אלא מקודדות כסופרפוזיציה של וקטורים רבים. השיטות הקלאסיות כמו PCA ו-SVD נכשלות, לא בגלל יישום לקוי אלא בגלל הנחות תאורטיות שגויות.

הכלי המרכזי כיום הוא Sparse Dictionary Learning ובעיקר Sparse Autoencoders. הרעיון הוא לאמן רשת קטנה ש"תפרש" את האקטיבציות של הרשת הגדולה באמצעות בסיס דליל של "תכונות". אלו הלייטנטים. אך בפועל, השיטה אמנם מוצאת כיוונים מעניינים, אך לא מסבירה איך החישוב עצמו מתבצע. הלייטנטים הם תמונה סטטית של "מה הופעל" ולא תיאור של האלגוריתם שמיושם.

יש גם בעיות מהותיות: הפער בין האקטיבציות האמיתיות לשיחזורן גדול. המידע הגאומטרי בין תכונות הולך לאיבוד. ההנחה שהכול לינארי רחוקה מלהיות נכונה. והגרוע מכל היא העובדה שאין בכלל תיאוריה פורמלית שמסבירה מהי "תכונה", איך היא נוצרת, ומה הופך אותה ליחידה בסיסית של הבנה.

מכאן עולה כיוון חדשני: אולי הדרך הנכונה היא לא לפרש מודלים אחרי שאומנו, אלא לבנות מודלים שאפשר להבין מראש. מודלים עם אקטיבציות דיסקרטיות, אכיפת מודולריות, פונקציות הפעלה דלילות כמו Top-k או SoLU, או מבנים כמו Mixture-of-Experts שמחלקים את החישוב לתת-מודולים ברורים. המטרה היא לייצר רשתות שנבנות "חתוכות מראש" עם פרשנות לא כניתוח מאוחר אלא כהנחת יסוד של האימון.

גם תיאור הפונקציה של רכיב בודד הוא משימה קשה. למשל דוגמאות שמפעילות אותו יכולות להיות מבלבלות. שיטות ייחוס מבוססות גרדיאנט בעייתיות תאורטית ופרקטית. סינתזת תכונות בניית קלט שמפעיל רכיב עלולה לייצר דימויים לא אינפורמטיביים. השיטות המבטיחות ביותר הן אינטרוונציות סיבתיות: שינוי של ערך פנימי, ובחינה של ההשפעה על ההתנהגות החיצונית. כאן נכנסים לתמונה גם steering כלומר החדרה של כיוון ספציפי למרחב האקטיבציות וגם שימוש logit lens כדי לפענח השפעה ישירה על תוצאות על אקטיבציות הרשת.

הבעיה הגדולה היא שהרבה מההסברים נשמעים משכנעים אך לא עומדים במבחן. הם לא חוזים קונטרפקטואלים(לא מצליחים לנבא מה היה קורה אילו משהו היה שונה בתוך המודל), לא עוזרים לאבחן כשלי מודל, לא מאפשרים תיקון או שיפור בפועל. לכן המחברים מציעים סט שלם של דרכי אימות: האם ההסבר חוזה התנהגות אחרי ablation? האם ניתן לבנות מודל קטן שמאפשר לבדוק אם ההסבר נכון? האם הלייטנטים ( ייצוגים פנימיים כמו תכונות או רכיבים חישוביים) מסייעים במשימות בטיחות כמו זיהוי תוכן מזיק? האם נוכל להשתמש בהסברים כדי לשנות את התנהגות המודל?

המאמר מציע גם ליצור "אורגניזמים מודליים" שהם מהווים רשתות קטנות סטנדרטיות, עם מבנה פתוח, שאפשר לאמן שוב ושוב ולבדוק עליהן שיטות פרשנות. כמו שהביולוגיה התקדמה דרך עבודה על תסיסנית, כך תחום זה זקוק לרפרנס קבוע. זהו כלי תשתיתי חסר כיום.

החלק האחרון של המאמר מבהיר שמכניזם אינו עניין טכני בלבד. הוא נוגע למדיניות, לניטור, לבטיחות, ולשאלות פילוסופיות: מה נחשב הסבר טוב? איך אפשר לחבר בין המבנים המיקרוסקופיים לתפקוד גלובלי? אילו עקרונות כלליים ניתן לחלץ מרשתות שלמדו לפתור בעיות טוב יותר מבני אדם?

בסיכום, מדובר במאמר לא מתבייש לומר את האמת: אין עדיין תיאוריה מספקת לפירוק רשתות. ההנחות הלינאריות שבריריות. התכונות לא חיות לבד אלא בארכיטקטורת על. הפרשנות חייבת לקשור מבנה לתפקוד. והדרך קדימה, אולי, עוברת לא דרך פענוח אלא דרך ״דזיין״ חדש של הרשתות…

https://arxiv.org/abs/2501.16496