⚡️🚀המאמר היומי של מייק 13.08.24: ⚡️🚀

Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2

בזמן האחרון התחלתי להתעניין בשיטות interpretability של מודלי שפה גדולים בעקבות כמה בלוגים מאוד מעניינים של אנטרופיק, OpenAI ולאחר מכן גוגל בנושא הזה. המטרה כאן היא לשפוך קצת אור על הקופסא השחורה שנקראת LLM - הרי אנחנו לא באמת מבינים איך הם עובדים ומה גורם להם לפלוט תשובה כזו אור אחרת לפרומפט שלנו.

אז המאמר הזה חוקר אחת השיטות המנסות להבין איך מודל שפה מייצג קונספטים סמנטיים שונים. המאמר עושה זאת דרך חקר של אקטיבציות הנוירונים בשכבותיהם השונות של מודלי שפה. עקב כך שיטה זו משויכת למשפחת שיטות המכונות mechanistic interpretability. הרעיון שהמאמר דן בו נקרא SAE או Sparse AutoEncoders.

אז מה הרעיון העיקרי ב- SAE? אנו מנסים להציג אקטיבציות של שכבה מסוימת של LLM על יד וקטור ארוך הרבה יותר מווקטור האקטיבציות אך מאוד דליל. כלומר וקטור n-ממדי של האקטיבציות אנו מייצגים (עם SAE) עם וקטור באורך M >> n אך בווקטור האורך הזה יש פחות מ- n איברים לא שווים לאפס (דלילות). SAE במקרה הזה פשוט מאוד: שכבה אחת לינארית עם אקטיבציה לא לינארית באנקודר (של SAE) ושכבה אחת של דקודר. המטרה כמובן לאמן את SAE כך שיהיה ניתן לשחזר את האקטיבציות המקוריות מייצוגם (אחרי האנקודר).

אבל למה זה בכלל חשוב ואיך זה קשור ל-interpretability של LLMs. הנחת מוצא של גישה זו (הבלוג של אנטרופיק מדבר על זה בהרחבה) שכל נוירון (או קבוצת נוירונים) בשכבה (מסוימת) הוא ״נדלק״ (מקבל ערכים) על כמה קונספטים לא קשורים (נגיד כלב, מכונה וערפל). כלומר הוא סוג של תערובת עבור כמה קונספטים. אז הייצוג המופק על ידי SAE הוא למעשה מהווה ייצוג של כל קונספט (disentangled). כלומר עבור כל קונספט המקודד קבוצות נוירונים שונות בוקטור הדליל הזה.

אז מה המאמר הזה עושה? הוא מנסה לאתר שכבות שבהם SAE מאומן עם שגיאת שחזור מינימלית (עם רגולריזציה מתאימה) כלומר הוא מנסה להבין איזו שכבה ב-LLM (וגם בשכבות הפנימיות של בלוקי הטרנספורמר) מקודדת הכי טוב את הקונספטים הסמנטיים.

בימים הקרובים עוד כמה סקירות בנושא המרתק הזה.

https://arxiv.org/abs/2408.05147