⚡🚀המאמר היומי של מייק 07.06.24: ⚡🚀

Scaling and evaluating sparse autoencoders?

המאמר הזה של openai ממשיך את הקו המחקרי של antropic (https://www.anthropic.com/news/mapping-mind-language-model) המנסה לראות איך ניתן למצוא של קונספטים (מסלולים וויזואלים) בתוך נוירונים של מודלי שפה מאומנים. המאמר של אנטרופיק בגדול טוען שיש נוירונים הנדלקים על קונספטים (נגיד גשר הזהב( מסוימים ויש כאלו שמהווים ערבוב של קונספטים.

אבל איך ניתן להציג את קונספט באמצעות וקטור? מתברר שניתן להציג כל קונספט באמצעות וקטור ארוך אך מאוד דליל(sparse). אז נוירונים המהווים ערבוב של קונספטים ניתן להציג בתור סכום משוקלל של וקטורים דלילים אלו אחרי שמטילים את הסכום על מרחב האמבדינג המקורי של הטרנספורמר.

הוקטורים הדלילים המתאימים לקונספטים ניתן להפיק באמצעות אימון של sparse autoencoder של שכבה אחת לכל כיוון כאשר הייצוג באמצע (אחרי האנקודר) הוא וקטור דליל: במהלך האימון לוקחים ממנו את k הרכיבים הגדולים ביותר - אחרי ReLu

ויש כמובן חוקי Scaling מעניינים לגבי הייצוגים האלו. מאמר מעניין.

מאמר: https://cdn.openai.com/papers/sparse-autoencoders.pdf