⚡️🚀המאמר היומי של מייק -04.11.24: ⚡️🚀

Refusal in Language Models Is Mediated by a Single Direction

מאמר מעניין החוקר איך ניתן לגרום למודל שפה לתת תשובות רצויות יותר ורצויות פחות. מתברר שאפשר לגרום למודל להסביר לנו איך מכינים הרואין או שודדים בנק ונמלטים מהעונש עם אם מזיזים פלט של שכבה אחת במודל שפה. וגם ניתן למנוע ממודל ״לא מרוסן״ לתת תשובות לא פוגעניות ולפעמים להימנע מלענות על שאלות מסוכנות אם מזיזים את הפלטים של כל השכבות של מודל, כל אחת עם וקטור r\_l כאשר l זו מספר השכבה.

איך בעצם מוצאים את הוקטורים האלה? עבור דאטהסט המכיל שאלות ותשובות רצויות מחשבים את ההפרש r הממוצע (על כל התשובות) בין האקטיבציות של כל שכבות המודל ועבור כל הטוקנים של חלון ההקשר. כלומר יש לנו מטריצה LxI של וקטורי ההפרש כאשר L זה מספר השכבות ו I זה מספר הטוקנים בחלון ההקשר.

כדי לגרום למודל להיות ״פחות מרוסן״ אנו בוחרים שכבה שהוספתן של מורידה ממנו את בלמים בצורה המשמעותית ביותר (יש מדדים לא רעים לכך). כלומר משאירים I וקטורי הפרשים שחישבנו. כדי לגרום למודל להיות יותר מנומס צריך להחסיר את ״כיוון הגסות״ מכל השכבות של המודל בצורה שתעביר אותם ממרחב אורתוגונלי ל r (כל שכבה ולכל טוקן בחלון ההקשר). בפרט מכל אקטיבציה x בכל שכבה ובכל טוקן : r \* r^T \*x קל לראות שהווקטור המתקבל כתוצאה מכך יהיה אורתוגונלי ל r.

עושים זאת לווקטור האקטיבציה לפני residual connection בכל בלוק של טרנספורמר. כמובן (מכיוון שיש הרבה מכפלות של מטריצות)ניתן להזיז גם את המשקלים שלהם כדי לקבל את אותם האפקטים. מאמר די מגניב וקל להבנה.

https://arxiv.org/abs/2406.11717