⚡️🚀המאמר היומי של מייק 15.07.24: ⚡️🚀

Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using Only Attention Maps

אהבתי את המאמר הזה כי הרעיון מאחריו הוא מאוד אינטואיטיבי ופשוט. המאמר מציע גישה להתמודדות עם הזיות(hallucinations) של מודלי שפה. מאוד בגדול הזיות של מודל שפה קורות כאשר מודל שפה עונה לא נכון לשאלת המשתמש. יש לכך כמה סיבות: למשל המודל לא מסוגל לענות על התשובה כי היא פשוט לא נמצאת ב״זכרון שלו״ (למשל השאלה על אירוע עדכני שהמודל לא אומן על הדאטה לגביו). הסיבה השניה היא העדר יכולות להבין את השאלה.

המחברים מנסים להתמודד עם הזיות של מודל שפה על ידי ניתוח של משקלי ה-attention של הפרומפט (השאלה) ושל תשובתו של המודל. נניח שהפרומפט מכיל N טוקנים וכרגע אנו חוזים טוקן מספר t של תשובתו של מודל שפה. קודם כל מחשבים את סכום מקדמי ה-attention עבור N טוקנים של הפרומפט P וסכום מקדמי ה-attention עבור כל t הטוקנים של התשובה R. מחשבים את היחס בין P ל- P + R ועבור כל שכבה של הטרנספורמר ועבור כל ראש (head) של בלוק הטרנספורמרים.

לאחר מכן בונים וקטור מהיחסים האלו ומאמנים מודל המכיל שכבה אחת שמטרתו היא לחזות האם המודל הוזה או לא. כיוון האורך תשובתו של המודל יכול להיות כלשהו המחברים מאמנים מודל עבור מספר קבוע של טוקני התשובה T. אם התשובה מכילה יותר מ- T וטוקנים מפעילים את המודל עבור כמה פעמים בשביל לזהות הזיות בחלקים השונים של התשובה.

איך בונים דאטהסט לאימון של המסווג הזה? בגדול נותנים למודל שפה לענות על שאלה ומפעילים מודל שפה חזק כדי לזהות תשובות נכונות ולא נכונות (הזיות).

https://arxiv.org/abs/2407.07071