⚡️🚀המאמר היומי של מייק -21.11.24: ⚡️🚀

Unfamiliar Finetuning Examples Control How Language Models Hallucinate

מאמר של סרגיי לווין האגדי מאוניברסיטת טורונטו שידוע יותר בתרומתו האדירה לפיתוח שיטות מבוססות למידה עם חיזוקים (RL) ליישומי רובוטיקה. הפעם הוא עם קבוצתו חוקר את תופעת הזיות (hallucinations) של מודלי שפה. הזיה זה מושג מאוד רחב בהקשר מודלי שפה ובגדול (מאוד) ניתן להגדירו בתור מתן תשובה לא נכונה (בעיקר עובדתית) על ידי מודל שפה.

מאז שמודלי שפה נכנסו לחיינו בשנים האחרונות תופעה זו נחקרה באופן נרחב בעשרות (אם לא מאות) מאמרים. המאמר שנסקור היום חוקר סיבות לתופעה זו וגם מציע דרכים להתמודד איתה. החוקרים טוענים הסיבה להזיות טמונה בניסיון להקנות למודל ידע חדש במהלך טיוב (finetuning). המחברים טוענים שהמודל נוטה ללמוד פחות טוב את העובדות הנמצאות בדאטהטס של FT (נקרא לו D\_FT) שלא מיוצגות מספיק טוב בדאטהסט הגדול ששימש את המודל לאימון מקדים (נקרא לדאטהסט זה בתור D\_PR). עובדות (ושאלות עליהם) נקראות לא-מוכרות במאמר.

בפרט המאמר משער (ומראה אמפירית) שעבור שאלה על עובדה לא q המודל מוציא תשובה שהיא סוג של תשובה ממוצעת עבור כל השאלות הלא מוכרות מ-D\_FT. כלומר

כזו שממזערת את פונקציית הלוס הממוצעת על כל השאלות הלא מוכרות האלו מ-D\_FT. ומכיוון שרוב התשובות ב-D\_FT מנוסחת היטב ובאנגלית רהוטה אנו מקבלים תשובות יפות אך לא נכונות בהחלט ממודל שפה לשאלות לא מוכרות.

בגדול הרעיון העיקרי שהמחברים מציעים לתיקון המצב הזה הוא ללמד את המודל להגיד ״לא יודע״ בצורה ברורה על שאלות לא מוכרות (כלומר במקרים שהוא אכן לא יודע). אחת הדרכים לעשות זאת היא קודם לזהות שאלות לא מיוצגות מספיק ב-D\_FT (על ידי ניתוח שכיחותם או אנטרופיה של הלוגיטים של תשובה המודל לשאלות אלו - ד״א שניהם לא אידאליים באספקט הזה). לאחר מכן במקום לאמן מודל לענות תשובות נכונות לשאלות אלו (שהוא לא מסוגל ללמוד), תשובות אלו מוחלפות ב-D\_FT על ידי תשובות נייטרליות בסגנון ״אני לא יודע״. כמובן אפשר להוסיף ל-D\_FT מלא שאלות הלא מוכרות ב-D\_PR עם תשובות אלו.

הדרך השנייה היא לאמן מודל עם שיטות של RLHF עם שינוי של פונקציית תגמול (reward) המקטין קנס על תשובות נייטרליות ומשאיר את שאר התגמולים כמו שהם. המחברים מראים (אמפירית) שבמקרה זה המודל יותר ״שמח״ לתת תשובות נייטרליות לשאלות לא מוכרות. המאמר מציע שיטה המורכבת מ-4 שלבים לאימון RLHF לשיפור יכולת המודל להגיד ״לא יודע״:

עושים FT רגיל

דוגמים את המודל עם שאלות מוכרות ולא מוכרות

בונים פונקצית תגמול הקונסת את המודל יותר על תשובות לא נכונות לשאלות לא מוכרות (וקנס מאוד נמוך או 0 על תשובות מתחמקות)

אימון RLHF עם פונקצית התגמול מסעיף 3.

מאמר נחמד שהשאיר בי טעם לראות את ההמשך.

https://arxiv.org/abs/2403.05612