⚡️🚀המאמר היומי של מייק 20.06.24:⚡️🚀

WARM: On the Benefits of Weight Averaged Reward Models

הסקירה הזו ממשיכה את קו הסקירות בנושא שיפור ביצועי RLHF לטיוב מודלי שפה. כבר דיברנו בסקירות הקודמות על כך שבמהלך RLHF המודל יכול לבצע reward hacking כלומר להתכנס לפוליסי (משקלי המודל) שממקסם את ה-reward ובאותו הזמן יוצר תשובות באיכות ירודה לפרומפטים.

המאמר שנסקור קצרות היום מציע לאמן כמה מודלי reward שונים ולהשתמש בממוצע שלהם כ-reward יותר ״יציב״ שעשוי למנוע מהמודל לעשות reward hacking. הבעיה העיקרית בגישה הזאת נובעת מכך שהיא מצריכה להחזיק בזמן אימון RLHF כמה מודלי reward שכמובן דורש יותר משאבי חישוב (ומייקר את חשבון החשמל).

המחברים מציע לשלב את התוצאה של המודלים אלא הביצועים שלהם. בשפה פשוטה הם מאמנים כמה מודלי reward וממצעים את המשקלים שלהם. זה מסתמך על איזושהי תופעה שלא ידעתי עליה שנקראת ״Linear mode connectivity״ או LMC הטוענת שהביצועים של מודל עם סכום ממושקל של המשקלים של כמה מודלים אחרים הוא יותר טוב מסכום ממושקל (עם אותם משקלים) של ביצועי המודלים (אולי אתעמק בזה בהמשך).

עכשיו כדי לבצע את הפעולה הזו הרשתות צריכות להיות בעלי אותה ארכיטקטורה ומה שונה בין מודלי reward כאן הם פרמטרי אימון כמו קצב למידה ודרופאאוט, סדר שונה של הכנסת דאטה לאימון (סיד שונה כנראה) וגם איתחולים שונים (לוקחים מודלים אחרי צ'קפוינטים שונים ב-SFT).

כתוצאה מקבלים מודל reward אחד טוב יותר שמשמש אותם לאימון RLHF.

https://arxiv.org/abs/2401.12187