⚡️🚀המאמר היומי של מייק 25.06.24:⚡️🚀

Improving Reinforcement Learning from Human Feedback with Efficient Reward Model Ensemble

הסקירה הזו ממשיכה את קו הסקירות על המאמרים שמנסים לשפר שיטות RLHF לטיוב (instruction tuning או פשוט fine-tuning) של מודלי שפה. בחלק של שיטת RLHF (למשל PPO) אנו מאמנים מודל reward מבוסס על סט של שאלות ותשובות מדורגות על ידי המתייגים האנושיים. מטרה של מודל זה לספק ציון לזוג (שאלה, תשובה) כאשר ציון גבוה מצביע על תשובה טובה ורצויה. לאחר כן אנו מאמנים (מטייבים) מודל שפה כאשר המטרה היא מקסום של פונקציה reward תוך שמירת של משקלי המודל למשקלים שהתחלנו מהם (נמדד על ידי KL divergence בין התפלגויות הטוקנים של שני המודלים). כל זה מתבצע on-the-fly כאשר הדוגמאות נוצרות עלי ידי הגרסה העדכנית של המודל במהלך האימון.

הבעיה עם הגישה היא reward hacking כאשר למרות איבר הרגולריזציה (KL) המודל מתכנס למשקלים שמגיעים לערכים גבוהים של פונקציית reward כאשר המודל עצמו ״לא מספק את הסחורה״. המאמר מציע להשתמש בכמה מודלי reward כי ensemble זה תמיד טוב. הבעיה שלהחזיק יותר ממודל אחד בזמן האימון זה יקר מבחינת המשאבים. המאמר מציע שתי גישות להתגבר על זה:

מתחילים מאותו המודל (שפה)

לאמן מודלי reward זהים עם ראשים לינאריים (מאומנים) שונים. כך צריך לשמור רק מודל אחד והמשקלים עבור השכבה הלינארית עבור כל מודל.

לאמן כמה מודלי reward בשיטה של LoRa - כך נצטרך לשמור רק את תוספת המשקלים לכל שכבה שזה יכול להיות די זול מבחינת המשאבים

ואז אפשר לקחת ממוצע של ה-rewards של כל המודלים או את המינומום ביניהם- יש לא מעט אופציות…

https://arxiv.org/abs/2401.16635