המאמר היומי של מייק - 24.01.25  
Back to Basics: Revisiting REINFORCE Style Optimization for Learning from Human Feedback in LLMs

תמצית המאמר:

המאמר בוחן מחדש את השימוש בלמידה מחיזוקים מפידבק אנושי (RLHF) באופטימיזצית LLMs. הוא מאתגר את הדומיננטיות של PPO (Proximal Policy Optimization) כשיטת למידת החיזוקים הסטנדרטית בהקשר זה, תוך הדגשת חוסר היעילות החישובית והמורכבות המיותרת שלו. במקום זאת, החוקרים מציעים לחזור לשיטות פשוטות יותר בסגנון REINFORCE, ספציפית Vanilla Policy Gradient (REINFORCE) והרחבתו הרב-דגימתית, REINFORCE Leave-One-Out (RLOO). שיטות אלו מוכיחות ביצועים טובים יותר מ-PPO מבחינת עלות חישובית, יעילות דגימה אופטימיזצית תגמול במספר מערכי נתונים וארכיטקטורות LLM. הממצאים מדגישים שניתן להשיג התאמה של LLMs להעדפות אנושיות עם אסטרטגיות אופטימיזציה פשוטות יותר המותאמות לייחודיות של RLHF.

הרחבה על נקודות עיקריות:

1. פישוט תיאורטי:

החוקרים מראים שרבים מהרכיבים של PPO (למשל, קליפינג, פונקציות ערך (value), ומידול ברמת טוקנים) אינם הכרחיים ל-RLHF, בהינתן האתחול טוב של LLM (לאחר SFT למשל). על ידי מידול סדרות שלמוץ כפעולות בודדות, REINFORCE נמנע מהמורכבות של פונקציות ערך-מצב(V ו-Q) ברמת טוקן, והופך את הבעיה לדומה יותר לבנדיט הקשרי.

2. יעילות מעשית:

שיטת RLOO משתמשת בכל הדגימות שנוצרו לבניית בסיס השוואה, משיג יעילות דגימה גבוהה יותר מ-RAFT, שלוקח את רק את הדגימות בעלות ציונים גבוהים (סוג של rejection sampling). זה מוביל לחיסכון משמעותי בחישובים וניצול טוב יותר של הנתונים הזמינים. הגישה מפשטת את תהליכי ה-RLHF על ידי הפחתת התלות בהיפר-פרמטרים רגישים כמו יחסי קליפינג והפרמטרים בשערוך פונקציית יתרון (כמו ב-GAE).

4.רובסטיות:  
שיטת RLOO מדגימה רובסטיות לתגמולים רועשים ועונשי KL גבוהים יותר, עולה על שיטות כמו RAFT שרגישות יותר לדיוקם.

תובנות תיאורטיות:

1. איזון שונות-הטיה(bias-variance tradeoff) ושערוך גרדיאנט ללא הטיה:

שיטת PPO מסתמכת על פונקציות ערך-מצב ושערוך יתרון מוכלל (Generalized Advantage Estimation) להפחתת שונות של שערוך גרדיאנט במחיר של העלאת הטיה. המאמר טוען שב-RLHF עבור LLMs די מאומן (warm start) הופך את הפחתת השונות לפחות קריטית. זה מאפשר לשיטות ללא הטיה כמו REINFORCE לתפקד היטב בלי להכניס הטיה משמעותית. אמפירית, המאמר מדגים ש-REINFORCE משיג אופטימיזציית תגמול טובה יותר מ-PPO, אפילו תחת תנאים של שונות גבוהה תיאורטית.

2. מידול מסלול מלא(תשובה שלמה) לעומת מידול ברמת טוקנים:

שיטת PPO ממדלת כל טוקן כפעולה, יוצר תהליך החלטה מרקובי (MDP) בו רצפי טוקנים חלקיים הם מצבים. עם זאת, RLHF מייחס תגמולים רק לתשובות שלמות, מה שהופך מצבי ביניים ללא רלוונטיים. על ידי מידול התשובה כפעולה יחידה, REINFORCE מפשט את הבעיה למבנה Contextual Bandit המתיישר ישירות עם מבנה התגמול. תוצאות אמפיריות מאשרות כי גישה זו עולה על מידול ברמת טוקנים הן ביעילות והן בביצועים.

3. קליפינג ויציבות עדכוני מדיניות:

שיטת PPO משתמש במנגנון קליפינג למניעת עדכוני פוליסי גדולים שעלולים לערער את הלמידה. החוקרים מראים שזה מיותר עבור RLHF, מכיוון שמשטח האופטימיזציה יציב הודות ל warm-started LLM. הסרת הקליפינג ב-PPO או הימנעות ממנו לחלוטין עם REINFORCE מובילה לביצועים טובים יותר, מה שמצביע על כך ש-RLHF אינו דורש רמה כזו של ייצוב.

4. איזון בין הפחתת שונות והעלאה קלה בהטיה:

אומדן היתרון ב- PPO מאזן בין שונות והטיה, נשלט על ידי ההיפר-פרמטר λ. ערכי λ גבוהים יותר (קרובים ל-1) מפחיתים הטיה אך מגדילים שונות. החוקרים מדגימים שב-RLHF, ערכי λ גבוהים יותר מובילים באופן עקבי לתגמולים טובים יותר של המודל, תומכים בשימוש באומדנים ללא הטיה כמו REINFORCE

מגבלות וכיוונים עתידיים

1. אופטימיזציית יתר של תגמול:

המחקר אינו מתמודד עם אופטימיזציית יתר של מודל התגמול(reward hacking), בה המדיניות מנצלת הטיות בפונקציית התגמול על חשבון הכללה. זה נשאר אתגר פתוח עבור RLHF.

2. הערכה אנושית:

בעוד שאחוזי ניצחון מדומים באמצעות GPT-4 משמשים כמדד להעדפות אנושיות, הערכות אנושיות ישירות היו מספקות ראיות חזקות יותר לאיכות ההתאמה.

3. "סקלביליות":

הסקלביליות של REINFORCE ו-RLOO למודלים(הם בדקו רק מודלים של 7B) ודאטהסטים גדולים יותר מצריכה מחקר נוסף.

מסקנה

המאמר מציג טיעון משכנע לבחינה מחודשת של שיטות בסגנון REINFORCE ב-RLHF, מאתגר את הדומיננטיות של PPO ודומיה. על ידי ניצול המאפיינים הספציפיים של RLHF - כמו warm started LLM ותגמולים ברמת הסדרה - החוקרים מדגימים ששיטות פשוטות יותר כמו REINFORCE ו-RLOO יכולות לעלות על חלופות מורכבות יותר כמו PPO ו-RAFT מבחינת אופטימיזציית תגמול, יעילות דגימה ועמידות.

https://arxiv.org/abs/2402.14740