המאמר היומי של מייק - 17.01.25  
Towards a Unified View of Preference Learning for Large Language Models: A Survey

מוטיבציה

המאמר מספק סקירה נרחבת של שלב מהותי באימון LLMs: יישור (alignment) של פלט המודל עם העדפות אנושיות. מיותר לציין כי יישור זה חיוני ליישומים רבים LLMs. בעוד ש RLHF וכיוונון מונחה (SFT) היו מרכזיים ליישור, היחסים ביניהם נותרו לא נחקרו מספיק, מה שמוביל לפיצול המאמצים המחקריים בנושאים אלו.

המחברים שואפים לאחד מאמצים מפוצלים אלה על ידי הצגת מסגרת המשלבת גישות RLHF ו-SFT תחת נוסחה מבוססת גרדיאנט אחת בלבד. איחוד זה לא רק מגשר על פערים מתודולוגיים אלא גם מכין את הקרקע להתקדמויות מגובשות יותר בלמידת העדפות (preference learning). המאמר מדגיש יישור כולל מספר מרכיבים - מודל, דאטה,משוב (כגון פונקציית תגמול עבור RLHF) ואלגוריתם - כל אחד הוא חשוב להבטחת (בתקווה) ביצועים חזקים.

תרומות טכניות:

נוסחת גרדיאנט מאוחדת לשני המקרי בלב המאמר נמצאת הנוסחה של גרדיאנט מאוחד לאופטימיזציה של העדפות (נוסחה 1 במאמר)

כאשר:

δ: מקדם גרדיאנט שתלוי באלגוריתם הספציפי, במשוב ובדאטה.  
A: האלגוריתם האופטימיזציה המיושם.  
r : אות משוב (feedback) המשפיע על מקדם הגרדיאנט (למשל תגמול)  
π\_θ: מודל מדיניות המפורמטר על ידי θ.

משוואה זו מכלילה את תהליכי האופטימיזציה המשמשים הן בשיטות מבוססות RL והן בשיטות מבוססות SFT, ומראה שההבדל העיקרי ביניהן טמון באופן שבו המשוב משולב. שיטות מבוססות RL משתמשות בדרך כלל בתגמולים סקלריים, בעוד ש-SFT משתמש בתוויות העדפה או דירוגים.

טקסונומיה של למידת העדפות:

המאמר מסווג למידת העדפות לארבעה שלבים מקושרים:  
דאטה:

דגימת דאטה On-Policy: דאטה נוצרים בזמן אמת על ידי המודל המאומן. טכניקות דגימה כמו Top-K, Nucleus Sampling, ו-Monte Carlo Tree Search משמשות ליצירת דאטה מגוון ואיכותי.

איסוף נתונים Off-Policy: הנתונים נאספים מראש, לעתים קרובות ממקורות חיצוניים, כולל סטי נתונים מתויגים על ידי בני אדם (כמו למשל בשיטות HH-RLHF, SHP) או דאטהסטים סינטטיים שנוצרו על ידי LLMs (למשל, UltraChat, ULTRAFEEDBACK).

משוב:

משוב ישיר: כולל תוויות אנושיות וחוקים המנוסחים על ידי בני אדם. דוגמאות כוללות בדיקות נכונות בחשיבה מתמטית או תוצאות יוניטסטים בייצור קוד.

משוב מבוסס מודל:

מודלי תגמול: מעריכים הסתברויות העדפה אנושית באמצעות שיטות כמו מודל Bradley-Terry (נוסחה 2 במאמר):

האופטימיזציה מושגת דרך פונקציית לוס סטנדרטית של לוג הנראות שלילית:

מודל תגמול מבוסס מסווג בינארי (למשימות בהן איכות המקרה ניתנת לקביעה על ידי תוצאותיו):

תיוג ישיר של דגימות לאימון מסווג בינארי כמודל תגמול היאגישה פשוטה ויציבה. למשל, בחשיבה מתמטית, ניתן לתייג  
דגימה על בסיס האם התשובה מניבה את התשובה הסופית הנכונה. באופן דומה, במשימות ייצור קוד, ניתן לבצע תיוג על ידי בדיקה האם הקוד שנוצר עובר בדיקות מוגדרות. בניגוד למשימות כמו סיכום טקסט או יצירת דיאלוג, הדורשות השוואות זוגיות של דוגמאות, שיטות הערכה ישירות אלו מפשטות את תהליך תיוג ההעדפות.  
  
בניגוד למודל התגמול המסורתי של Bradley-Terry, ברגע שיש לנו התיוגים עבור הדאטה, ניתן לאמן את מודל התגמול באמצעות פונקציית לוס של סיווג בינארי תגמול עבור כל לייבל מבלי שיהיה צורך לבנות דאטה עבור זוגות.

שיטת LLM-as-a-judge: משתמש ב-LLMs עצמם להערכת פלטים. מנגנוני תגמול עצמי, מטא-תגמול (מודל שפה בונה ציון עבור ה-verdict שהוא בעצמו נותן) ועוד מגוון שיטות בסגנון.

אלגוריתמים:

האלגוריתמים מחולקים לקבוצות על פי מספר הדגימות הנדרשות לחישוב הגרדיאנט:

שיטות Point-Wise: אופטימיזציה באמצעות דגימות בודדות. דוגמאות כוללות (Proximal Policy Optimization (PPO ו-ReMax.

שיטות Pair-Wise Contrast (סוג של למידה ניגודית): מנצלות השוואות בין זוגות של דגימות: הדוגמא הבולטת של שיטה זו היא (Direct Preference Optimization (DPO.

שיטות List-Wise Contrast: משערכות את הגרדיאנט על פני כמה דגימות. גישה זו שימושית במיוחד במשימות הדורשות הערכות הוליסטיות, כמו דירוג או סיכום.

שיטות Training-Free: כוללות טכניקות אופטימיזציה של קלט/פלט, המבטלות את הצורך בעדכוני גרדיאנט במהלך היישור (המאמר לא מרחיב על זה)

אבלואציה:

אסטרטגיות אבלואציה בוחנות עד כמה טוב LLMs מתיישרים עם העדפות אנושיות.

הערכה מבוססת חוקים: משתמשת בקריטריונים מוגדרים מראש כמו נכונות עובדתית או אמות מידה ספציפיות למשימה.

הערכה מבוססת LLM: כוללת LLMs מתקדמים הפועלים כמעריכים, משתמשים בפרומפטים להערכה ודירוג תגובות.

טכניקות דגימת דאטה:

דגימה On-Policy: MCTS משפר את עושר של הדאטה איכותו למשימות הדורשות חשיבה רב-שלבית.

דגימה Off-Policy: דאטהסטים סינתטיים, המיוצרים על ידי LLMs מתקדמיםֿ, משמשים יותר ויותר כדי ״לתת סקייל״ ללמידת העדפות.

מסקנה:

סקירה זו מספקת מבט מתמטי קפדני ומאוחד מושגית על למידת העדפות עבור LLMs. המסגרת שלה מבהירה יחסים בין שיטות RL ו-SFT, מאפשרת לחוקרים להשוות, לשלב ולחדש אסטרטגיות יישור העדפות באופן שיטתי. הדגש על משוב, תכנון אלגוריתמים והערכה מבטיח כיסוי מקיף של התחום, הופך מאמר זה למשאב יקר ערך לקידום מחקר יישור LLM.

https://arxiv.org/abs/2409.02795