המאמר היומי של מייק: 05.09.25

Group Sequence Policy Optimization

סקירה מס' 500:

סקירה מס' 500 והדי חגיגית לכאורה, בהתחלה חשבתי לבחור איזה מאמר מיוחד אבל לאחר הרהורים עמוקים (אך לא ארוכים) החלטתי לדחות את החגיגה למאמר מס' 512. ושם כבר נחליט, אולי נדחה ל-555 או משהו כזה - נראה איך התקדמו ההפתעות ששותפיי ואני מכינים לכם 🙂.

המאמר המציע שכלול לשיטת GRPO או Group Relative Policy Optimization השייכת למשפחת שיטות RHLF המשמשות לאימון ולפיינטיון של מודלי שפה. השיטה המוצעת שקיבלה שם GSPO (החליפו Relative ב- Sequence) במאמר משנה את פונקציית המטרה של GRPO.

בגדול מאוד GRPO ממקסם את המכפלה של שני הגורמים(יש גם כמה פונקציות קליפ שם). הגורם הראשון הוא מה הוא היתרון של הפוליסי הנוכחי (שזה בעצם ההתפלגות המותנית של הטוקן בהינתן ההקשר הקודם לו) על הפוליסי הישן (שממנו נגדמים הטוקנים באימון). GRPO להבדיל מ-PPO הקלאסי לא מחשב אותה דרך פונקציית value אלא מחשבת אותו יחסית לתגמולים (rewards) יחסית לתגמולים המתקבלים עבור הטוקנים הנדגמים עבור אותו הפרומפט (בגלל זה מילה group מופיעה בשם של השיטה).

הגורם השני הוא היחס של פולוסי החדש שאותו למעשה מאפטמים (ההתפלגות המותנית של טוקני המודל) לפוליסי הישן שממנו נגדמים הטוקנים. כאן בא ההבדל העיקרי בין GRPO לשיטה המוצעת נמצא באיך מחשבים את היחס הזה. ב-GRPO מחשבים את זה בתור יחס של הפוליסי החדש והישן ברמת הטוקן מנורמלי באורך תשובה עד הטוקן הזה. חישוב זה כמובן בעל שונות גבוהה וזה הסיבה להימצאות בפונקציית המטרה שם כמה קליפים כדי למנוע שינויים גדולים יותר. ד״א ב-PPO החישוב מתבצע ברמת התשובה כולה אבל שהופך את התגמולים לדלילים (sparse) שזה כמובן תרחיש לא פשוט בבעיות RL.

המאמר מציע 2 שיטות. הראשונה, שמחזירה את החישוב ברמת התשובה כולה, מחשבת את היחס בתור ממוצע על ההסתברויות של כל טוקנים (בלוג סקייל) כאשר כל אחת מהן מנורמלת באורך של התשובה עד הטוקן הזה. השיטה המוצעת השניה משאירה, בדומה ל-GRPO, את החישוב ברמת הטוקן אבל יחס ההסתברות עבור כל טוקן מחושב בצורה דומה לשיטה הראשונה - רק שהממוצע מחושב על הטוקן. שתי השיטות נראות בעלות שונות קטנה יותר מ- GRPO אבל הקליפים עדיין נמצאים בפונקצייה המטרה.

יש במאמר לא מעט טענות לגבי הקשר בין השיטה המוצעת ופונקציית המטרה של GRPO ו-PPO ל-importance sampling או IM. אזכיר כי IM היא שיטת דגימה מהתפלגות P שקשה לדגום ממנו באמצעות דגימה מהתפלגות Q שיותר קל לדגום ממנה. משקל ה-importance עבור דגימה x הוא יחס של ההסתברות של x עם P ועם Q. אמנם יש קשר אמיתי בין IM לשיטות המוזכרות במאמר אני לא השתכנעתי שכל הנימוקים במאמר הם נכונים מתמטית - יש מצב שאני לא הבנתי אותם מספיק עמוק.

כך או כך מאמר מעניין וראוי להיות מספר 500!

https://www.arxiv.org/abs/2507.18071

האות והרעש: פיזיקה חדשה להערכת מודלי שפה