⚡️🚀המאמר היומי של מייק -18.11.24: ⚡️🚀

Knowledge Editing in Language Models via Adapted Direct Preference Optimization

היום סוקרים מאמר כחול לבן בנושא פיינטיון(=טיוב, ככה אמרו לי) של מודלי שפה באמצעות טכניקות מבוססות על למידה עם חיזוקים או בקצרה RLHF. למיטב ידיעתי השימוש הראשון ב-RLHF היה במאמר במאמר שפיתח מודל הנקרא InstructGPT שמשמו כבר ברור כי אומן לעקוב אחרי הוראות המשתמשים. זה נעשה באמצעות טכניקת RL הנקראת Proximal Preference Optimization או PPO. מעניין ש-PPO הומצאה על ידי לא אחר אלא ג'ו שולמן שהיה תקופה די ארוכה CTO של OpenAI. בגדול מאמנים את המודל למקסם את פונקצית התגמול של תשובותיו תוך שמירתו (התפלגות הטוקנים) קרוב יחסית להתפלגות ההתחלתית (דרך KL divergence).

החיסרון העיקרי של PPO היה צורך באימון מודל תגמול (reward) שבהינתן שאלה ותשובה נותן ציון המשקף את איכות התשובה מנקודת ראיה של בני אדם (לפחות אלו שמאמנים מודלי שפה). לשמחתנו זמן קצר לאחר מכן הוצעה גישה הנקרא DPO או Direct PO שאפשרה לטייב (או ליישר כמו align) מודלי שפה ללא צורך בלהשתמש במודל תגמול בצורה מפורשת (מניחים צורה אופטימלית של התגמול נפטרים ממנו). כדי לאמן מודל שפה בשיטת DPO צריך דאטהסט המורכב מתשובות רצויות יותר ורצויות פחות ואנו מאמנים מודל.

המאמר למעשה פיתח שיטה שהתאימה את DPO לבעיה של עריכת ידע (knowledge editing) של מודל שפה. כלומר אנו רוצים שהמודל יענה אחרת על שאלות מסוימות (נגיד מתאימים אותו לדומיין מסוים). בעיה זו שקולה לבעיית יישור מודל שפה שניתן לפתור עם DPO. המחברים הציעו 3 שכלולים עיקריים ל-DPO:

במקום סט שאלות ותשובות(חיובית ושלילית) התשובות השליליות נוצרות על ידי המודל במהלך האימון

התשובות השליליות(המודל הנוכחי) מג'ונרטות עם מה שנקרא teacher-forcing. כלומר עד טוקן שחוזים משתמשים בטוקנים של התשובה החיובית (שאותה אנו מצפים לקבל מהמודל לאחר עריכת ידע)

האופטימיזציה עם DPO מבוצעת עם ה-teacher forcing הזה (נשמע מאוד הגיוני עם 2)

ויש תוצאות לא רעות כמובן…

https://arxiv.org/abs/2406.09920v1