⚡️🚀המאמר היומי של מייק 16.06.24:⚡️🚀

STATISTICAL REJECTION SAMPLING IMPROVES PREFERENCE OPTIMIZATION

המאמר הזה וכמה הבאים שאסקור בימים הקרובים מציעים שכלולים שונים לשיטה Direct Preference Optimization או בקיצור DPO. למעשה DPO בעצמה היא שדרוג של Proximal Policy Optimization או PPO שהפכה להיות מאוד פופולרית אחרי שמכמה חברות השתמשו בה ליישור מודלי שפה (alignment או instruction tuning) בתור השלב האחרון של אימון מודל שפה foundational. השיטה שייכת למשפחת RLHF כי היא דורשת דאטה (שאלות ותשובות) המדורגות על ידי בני אדם - עבור כל שאלה הם (המתייגים) בוחרים מה התשובה איזה תשובה טובה יותר.

למעשה DPO בא לייתר את מודל התגמול (reward) גם חוסך גם משאבים לאימונו וגם מאפשר לא להחזיק מודל נוסף בשלב RLHF. למעשה DPO מנצל את המבנה של פונקצית לוס של PPO, שהיא מקסום פונקציית תגמול עם איבר רגולריזציה שבא לשמור את המודל המיושר קרוב למודל התחלתי, כדי להיפטר מפונקציית התגמול בפונקציית לוס. זה מתאפשר עקב העובדה שקיים ביטוי מפורש לפוליסי האופטימלי (מודל שפה ״מושלם אחרי היישור") דרך הפוליסי אחרי ה-SFT (מודל שפה שאנו מתחילים ממנו את אימון היישור) ופונקציית התגמול.

אחרי שמשתמשים במודל לוס המושרה על ידי מודל (Bradley-Terry (BT המגדיר מהי הסתברות העדפה של תשובה חיובית על תשובה שלילית (על אותה השאלה) מה- rewards שלהם, ואנו מגיעים לביטוי עבור לוס של RLHF שמכיל רק את הפוליסי התחלתי. זה למעשה DPO והוא ממזער את פונקציית הלוס שלו על סט המכיל זוגות של תשובות טובות וגרועות.

המאמר שנסקור היום שואל את השאלה האם הדגימה האחידה מהסט הזה היא אופטימלית (מבחינת איכות התוצאה שהיא הופליסי הסופי או מודל שפה אחרי היישור). אולי אם היה לנו פונקציית תגמול היינו מעדיפים זוגות עם יחס מקסימלי בין ה-reward של התשובה החיובית לשלולית? אולי צריך לתעדף זוגות עם reward שלילי הנמוך ביותר?

המאמר מציע את הגישה הבאה:

- מאמנים מודל text2text שבהינתן שאלה ושתי תשובות מוציא את התשובה המועדפת.

- בעזרת המודל הזה בונים את פונקציית התגמול דרך סמלוץ (על ידי דגימה של שאלה וזוג תשובות) של הסתברות העדפה של תשובה טובה על תשובה גרועה.

- בעזרת פונקציית תגמול זו בונים פוליסי pi\_r שלמעשה זה מודל שפה (המאפשר לחשב הסתברות של תשובה בהינתן שאלה)

- משתמשים בדגימת rejection כדי לדגום pi\_r באמצעות הפוליסי ההתחלתי (= מודל שפה) כדי למזער את הלוס בדרך לפוליסי "המיושר".

הם גם משחקים עם כמה פונקציות לוס כמו hinge loss (בטח כבר שכחתם אבל אוהבים להשתמש בו ב -SVM).

https://arxiv.org/abs/2309.06657