המאמר היומי של מייק - 01.01.25:  
Inference-Aware Fine-Tuning for Best-of-N Sampling in Large Language Models

מתחילים את השנה החדשה עם סקירה של מאמר די מעניין שמציע שיטה לשיפור אימון של מודלי שפה. היתרון הגדול של השיטה היא מאפשרת להתאים את האימון לאופן ההיסק (אינפרנס) ודי ברור שאם אכן עושים זאת בהצלחה זה אמור להניב איכות ההיסק. כלומר אם אנו משתמשים בגישה מסוימת במהלך האינפרנס: למשל לבחור את ״התשובה הטובה ביותר״ מבין N תשובות המודל (המאמר מפתח שיטות רק לגישה זו וקורא לה BoN) או תיקון עצמי (self-correction) אז כדאי לנו להתאים את האינפרנס לכך.

קודם כל המאמר מנסח שתי פונקצית יעד לאימון inference-aware או IA בקצרה, אחת ל SFT וגם ל-RLHF, שקיבלו שמות IA-SFT ו-IA-RL בהתאמה. עבור IA-SFT אנו ממקסמים את נראות של תשובות המומחים לשאלות מהדאטהסט שיש לנו בהינתן פוליסי האינפרנס I (שזה למעשה BoN). למעשה מאפטמים את הפוליסי (שזה מנגנון חיזוי של LLM או בפשטות LLM עצמו) כדי לעשות את BoN בצורה הטובה ביותר על הדאטהסט שיש לנו. עבור IA-RL המטרה היא לאפטם את הפוליסי (שזה LLM כאמור) תחת טכניקה של אינפרנס I (כלומר BoN) כך שהיא ימקסם את פונקצית תגמול R.

לאחר מכן המאמר מגדיר באופן מדויק מה זה BoN (נוסחה 1) כאשר המטרה היא למקסם את איכות התשובות של המודל כאשר אנו בוחרים את התשובה לפי מה שנקרא verifier score (= ציון לאיכות התשובה). דרך אגב מתווספים כאן שני פרמטרים נוספים שהם מספר התשובות שמהן בוחרים את התשובה הכי טובה וגם טמפרטורת T של מודל השפה. באופן אינטואיטיבי ככל ש T גבוהה יותר (יותר רנדומליות ויצירתיות התשובות) מספר התשובות N צריך לעלות.

כאן יש לנו כאן הטרייד-אוף הקלאסי בין exploration ל-exploitation. ככל ש- T גבוה יותר אנו מבצעים יותר exploration (תשובות מגוונות יותר) ואילו T קטן יותר מאפשר לנו ״להנות״ ממה שלמדנו עד עכשיו (בחירה של N משפיע על הטרייד-אוף באופן הפוך מ-T).

אוקיי, אבל עדיין בבעיית אופטימיזציה של IA-SFT יש לנו את argmax (משתמשים בו לבחירה של התשובה הטובה ביותר) וזה מאוד מקשה על פתרונה למרות שיש לנו שיטות שערוך argmax באמצעות softmax וגם Gumbel softmax עדיין שיטות אלו אינן מדויקות וגם כבדות חישובית (לטענת המאמר). אז המחברים משתמשים בטריק מאוד מפורסם ב-ML - קירוב של פונקציית יעד עם קירוב וריאציוני שהופך אותה (את הלוג שלה) לסכום של הפוליסי (הסתברות של תשובה y בהינתן שאלה x עם המודל) ושל איבר רגולריזציה הנקרא inference aware. איבר זה הוא למעשה הוא win-rate של תשובה y על שאלה x על המודל הנוכחי כאשר הערך של כל זוג (x, y) מחושב עם verifier score r (עם קבוע נרמול).

ב- IA-RL הסיפור קצת מסתבך והמחברים משתמשים בתוצאה מאחד המאמרים של ג'ו שולמן (cto לשעבר של openai) עם סרגיי לווין ופיטר אבל האגדיים כדי לקבל שערוך לגרדיאנט שמקבל צורה דומה לאלגוריתם הישן והידוע REINFORCE כלומר המכפלה של הלוג של הפוליסי עם פונקציית תגמול (״ממורכזת״ עם התוחלת של פונקציית תגמול להקטנת השונות). המאמר גם דן במקרים מעניינים של אופטימיזציה של פונקצית היעד של IA-RL לכמה צורות של verifier score r (למשל בינארי).

מאמר די כבד מתמטית ניסיתי (לפי מיטב יכולתי) להנגיש לכם אותו טיפה…

https://arxiv.org/abs/2412.15287