⚡️🚀המאמר היומי של מייק 28.06.24:⚡️🚀

On-Policy Distillation OF LANGUAGE MODELS: LEARNING FROM SELF-GENERATED MISTAKES

מזמן לא סקרתי מאמר על שיטות זיקוק של ידע(knowledge distillation) - לא נתקלתי במאמרים מגניבים בנושא המעניין הזה. מה זה זיקוק ידע ממודל גדול למודל קטן יותר? למעשה זה ניסיון להעתיק למודל הקטן את הידע שיש למודל הגדול כלומר לגרום לו להפגין ביצועים הדומים למודל הגדול.

יש כמה שיטות לעשות זאת - הפשוטה ביותר זה לאמן אותו על הדאטה שהמודל הגדול אומן עליה. יש שיטות המאמנות את המודל הקטן על הדאטה המיוצר על ידי המודל הגדול. אם יש לנו גישה להתפלגויות (של הטוקנים) אז מאמנים את המודל הקטן לחקות את התפלגות הטוקנים שהמודל הגדול מוציא. אם יש לנו אקטיבציות של השכבות של המודל הגדול ניתן לנסות לחקות גם אותם (אם המודל הקטן הוא בעל אותה ארכיטקטורה אבל עם פחות שכבות).

בכל גישות האלו אנו מאמנים (או פיינטיון) את המודל הקטן בצורה supervised רגילה. כלומר יש לנו סט של דוגמאות (ground-truth או שנוצרו על ידי המודל הגדול) אנו מאמנים את המודל הקטן עליהם. המאמר שנסקור היום מציעה להשתמש בגישה מעולמות למידה באמצעות חיזוקים (reinforcement learning) ממשפחת on-policy. זה אומר שהאימון מתבצע על הדוגמאות שהרשת המאומנת עצמה יוצרת במהלך האימון (והיא משתנה כמובן).

המאמר הלך צעד אחד קדימה והחליט לשלב את שיטת אימון on-policy יחד עם האימון הסטנדרטי של זיקוק ידע. כלומר בהסתברות alpha השיטה בוחרת דוגמא מדאטהסט האימון ובשאר המקרים היא מגרילה דאטה מהמודל הקטן. כל פעם המודל מנסה למזער את המרחק בין התפלגות הטוקנים של הדוגמא (מהדאטהסט או מהמודל הקטן).

בד״כ כלל המרחק בין התפלגויות של הטוקנים בשיטות זיקוק ידע נמדד על KL divergence סטנדרטי (כלומר forward). המאמר מציע לשכלל את הגישה הזו עקב חולשה שיש ל- forward KL. החולשה הזו קשורה לעובדה ש-forward KL מנסה לקרב את התפלגות המודל המאומן לאזור המוד(mode) של התפלגות היעד (התפלגות המודל הגדול במקרה שלנו. הכוונה כאן שהתפלגות המודל המאומן עלולה ״להתרכז באזור בעל מסה הסתברותיות גבוהה״, נגיד ליד איזה מוד של ההתפלגות ומתעלמת מאיזורים אחרים שיש בהם מסה הסתברותית ליד מודים חלשים יותר של ההתפלגות.

למזלנו יש לנו reverse KL שהופך את המונה ואת המכנה בלוג של forward KL. ניתן להראות כי forward KL מנסה ״לכסות״ את כל האזור בה התפלגות היעד גדולה מאפס ובכך משלימה את forward KL. ניתן לשלב אותם לינארית (באופן קמור עם מקדם beta ו- 1-beta) ואז מקבל Jensen Shannon Convergence או JSD שנותן מענה לבעיה האינהרנטית של forward KL. ובה המאמר משתמש במקום forward KL הרגיל.

ניתן לשלב את פונקציית הלוס של המאמר עם עוד איבר האחראי על מקסום פונקציית reward כלשהי עבור המודל הקטן (כמו ב-RLHF).

ושכחתי להגיד(לא קשור למאמר) ש- forward KL זה בדיוק מה יש לנו בכל פונקציית לוס המבוססת על cross entropy (נגיד במשימות סיווג).

https://arxiv.org/abs/2306.13649