⚡️🚀המאמר היומי של מייק -16.10.24: ⚡️🚀  
EFFICIENT REINFORCEMENT LEARNING WITH LARGE LANGUAGE MODEL PRIORS

היום נסקור מאמר שהוא נראה די כבד מתמטית (הרבה נוסחאות ומלל שנראה מתמטי) אבל הרעיון מאחוריו הוא די פשוט וקל להסבר. אנחנו אוהבים למנף את עוצמתם של מודלי שפה למשימות רבות (ולא תמיד לכאלו שהם מסוגלים לבצע כמו שצריך לפחות כרגע).

המאמר מציע להשתמש במודל שפה כפריור עבור סוכנים במשימות בהם הם צריכים לבצע SDM או sequential decision making. המאמר נותן בתור דוגמא משחק overcooked כאשר הסוכן צריך לבצע משימות בישול שונות בהתבסס על מצב המטבח שבו הוא מבשל אותם. המטרה של הסוכן היא לחזות את הפעולה הבא (באמצעות תיאור טקסטואלי) כאשר התגמול הוא ביצוע נכון של המשימה (הכנה של מנה לפי המתכון :)).

כאמור המטרה כאן היא לחזות את הפעולה הבאה עבור הסוכן (המתוארת) על ידי הטקסט כאשר המצב (state) גם מתואר על ידי טקסט. בגדול מאוד אנו מתחילים ממודל אחד (הפריור P) עבור חיזוי המצב הבא (מהמצב הקודם והפעולה) ועבור חיזוי הפעולה הבאה בהינתן המצב (מתואר על ידי התפלגות Q\_h). המטרה כאן היא ללמוד את Q\_h כאשר ממקסמת התגמול הצפוי ושומרת את התפלגות Q קרובה לפריור P (זוכרים PPO שהתפרסם מאוד לפני שנתיים כאשר OpenAI השתמשו בו ל-RLHF לאימון מודלי שפה). המרחק כמובן ניתן על ידי ה-KL 🙂

אז הפעולה הבאה a\_t (כלומר גנרוט התיאור הטקסטואלי שלה) מתבצע באופן הבא. דוגמים כמה גרסאות של a\_t עם P מחשבים את הנראות שלהם לפי Q הנלמד, מנרמלים עם הסופטמקס ודוגמים את הפעולה הבאה כאשר מטרת התהליך מקסום של התגמול הצפוי (עם הרגולריזציה שהסברנו עליה קודם).

כמובן שניתן לעשות את זה בכמה אופנים: בצורה של online דרך שערוך של פונקציית Q של הזוג (מצב, פעולה) כאשר פונקציית Q קשורה להתפלגות Q\_h של הפעולה הבא שנידונה בפסקה הקודמת (עניין של נרמול נכון). ניתן לעשות את זה גם באמצעות offline עם איזה פוליסי טוב ידוע של המומחים כאשר המטרה היא גם שערוך של פונקציית Q שבאמצעותה ניתן לשערך (לקבל) את Q\_h עבור חיזוי הפעולה הבא. ניתן לעשות את זה גם באמצעות שיטה דומה ל-PPO אבל בכל המקרים הפריור הוא ההתפלגות המושרית על ידי מודל שפה נתון.

מאמר מעניין בקיצור…

https://arxiv.org/pdf/2410.07927