⚡️🚀המאמר היומי של מייק 08.09.24: ⚡️🚀

DO TRANSFORMER WORLD MODELS GIVE BETTER POLICY GRADIENTS?

לא הייתי אמור לכתוב סקירה היום אך הקלטת הפודקאסט שלנו התבטלה והתפנה לי קצת זמן אז אסקור מאמר שכבר נמצא כמה זמן אצלי במגירה. המאמר בנושא למידה עם חיזוקים (RL) וטרנספורמרים אז לכאורה זה נשמע מאמר די נחמד.

המאמר מדבר על שיטה לשיפור של למידה פוליסי בבעיות של RL בבעיות שיש לנו גישה ישירה לדינמיקה של הסביבה (כלומר אנו לא יכולים לאסוף עליה דאטה רלוונטי המאפיין את פיצ'רים המהותיים שלו קרי non-observable). בגדול המטרה שלנו בלמידת פוליסי היא לחזות את הפעולה (action) האופטימלי בהינתן המצב s של הסביבה והפעולה האחרונה s. אופטימלי כאן משמעותו מקסום של התגמול (reward) הכולל המתקבל במהלך אפיזודה. המודל שחוזה את הפעולה הזו הוא למעשה מממש את הפוליסי שלנו.

אבל מה לעשות אם אין לנו גישה ישירה לסביבה? במקרה הזה אנו יכולים לאמן מודל שהוא חוזה לנו את המצב הבא s בהינתן המצב הקודם(כלומר ייצוגו) והפעולה האחרונה(עם הנחת המרקוביות) או בהינתן N ייצוגים של המצבים האחרונים והפעולה האחרונה. זה למעשה נקרא world model (לדעתי יחד עם מודלים המשערכים את התגמול הצפוי למצב נתון - value function אבל זה פחות חשוב כרגע).

איך המודל הזה מאומן? מאינטראקציה עם הסביבה - הסוכן מבצע פעולות בסביבה ואנו מעדכנים את ה-world model שלנו בהתבסס על משוואות Bellman). שימו לב אם או ללא הנחת מקרוביות אנחנו משערכים את הייצוג של המצב ה״עולם״ הבא בהינתן המצב(-ים) הקודמים. המאמר טוען שזה יוצר גרדיאנטים לא יציבים ושונות גבוהה עקב שימוש ישיר בשערוך של המצבים הקודמים לשעורך של המצב הבא.

הם מציע לשערך את המצב הבא מהפעולה ולא מייצוגי המצבים שטענתם ״הופך את הגרדיאנטים במודל העולם לפחות מעגליים״ וזה תורם ליציבות השערוך. יש גם קצת הוכחות במאמר (סוג של) של הטענה הזו. המאמר מראה אם יש לנו מקרוביות (התלות של המצב הבא היא רק במצב האחרון) השיטה המוצעת עובדת כמו RNN מבחינת הגרדיאנטים. בתחושה זה נשמע לי די טבעי (אשמח אם מישהו ירחיב על זה). במקרה שאין לנו מרקוביות הטענה לביצועים טובים יותר של השיטה המוצעת.

לא ראיתי אזכור משמעותי מדי של הטרנספורמרים במאמר (תקנו אותי אם אני טועה).

https://arxiv.org/abs/2402.05290