⚡️🚀המאמר היומי של מייק 23.07.24: ⚡️🚀

Feedback Efficient Online Fine-Tuning of Diffusion Models

ממשיכים את הקו של אתמול וסוקרים עוד מאמר המשלב מודלי דיפוזיה עם טכניקות מעולם של למידה עם חיזוקים (RL). הפעם המאמר משלב את שני התחומים המרתקים האלו כדי לבצע פיין טיון של מודל דיפוזיה. המאמר מתמקד במקרה שאין בידינו דאטהסט (לפיין טיון) אלא יש לנו דרך לשערך (סוג של reward) את איכות של פיסת דאטה מג'ונרט, כלומר סוג של משוב על איכות הדאטה. למשל אם מטרתנו היא לאמן מודל לגנרט מולקולות המשוב יכול להיות ״מידת פעילות ביולוגית״ (bioactivity) של המולקולה הנוצרת.

בגדול מאוד המאמר מציע לאמן מודל דיפוזיה מאומן (pretrained) למקסום של פונקציית התגמול (=המשוב) תוך כדי שמירת של התפלגות הדאטה המגונרט על ידי המודל קרוב יחסית לזו של המודל ההתחלתי. מזכיר לכם PPO ו-TRPO מעולם ה-RL - אז זה בערך אותו הרעיון עם קצת סיבוכים. התהליך הוא איטרטיבי וכל איטרציה אנו מעדכנים את פרמטרי המודל (כאן זה רק המשקלים - יוסבר בהמשך) ויוצרים דאטה חדש עם המודל המעודכן.

למעשה התהליך מורכב מ 3 שלבים עיקריים.

בשלב הראשון בונים דאטהסט חדש עם מודל דיפוזיה מהאיטרציה הקודמת (בהתחלה מתחילים ממודל מאומן (pretrained). כמו שכתבתי ניתן לתאר מודל דיפוזיה מאומן על ידי משוואה דיפרנציאלית סטוכסטית עם אופיינים נלמדים (פונקציה נלמדת למעשה עם שיטות כמו score matching או flow matching). למעשה ה-SDE הזה מתאר את תהליך יצירת דאטה מרעש טהור. אז בשלב הראשון מתחילים מרעש מחדש ופותרים את ה-SDE (עם פונקציה נלמדת התלויה בדאטה מורעש באיטרציה t וב- t עצמו). משתמשים בשיטות סטנדרטיות כמו אוילר או אוילר מוריאמה.

בשלב השני בהתבסס על הדאטה שיצרנו בשלב הקודם מאמנים מודל מאמנים מודל תגמול reward עם רגולריזציה (נגיד L1 או כל פונקציה התלויה במאפייני ה-reward ובמשימה עצמה). בנוסף מאמנים מודל המשערך אי וודאות של פונקציית תגמול. בגדול במקרה הזה המטרה של הפונקציה היא שערוך של סוג של רווח סמך של הפרש של פונקציית התגמול אופטימלית עם רגולריזציה ופונקצית תגמול עצמה על הדאטהסט מהאיטרציה הקודמת(הפרטים קצת מורכבים והעדפתי לא לצלול בהם בסקירה).

בשלב השלישי של כל איטרציה מאמנים פונקציה חדשה f עבור ה-SDE שלנו וגם ההתפלגות ההתחלתית v של שממנה אנו מייצרים את הדאטה באמצעות ה-SDE. יש שם נוסחאות די מורכבות אך אנסה להסביר את ההיגיון מאחוריהם בכל זאת. פונקציית המטרה כאן מורכבת מ 3 איברים (ממקסמים אותה על הדאטהסט משלב 1). המקסום מתבצע ביחס לפונקציית f וגם על ההתפלגות ההתחלתית ממנה יוצרים את הדאטה באמצעות SDE:

התגמול האופטימיסטי (סכום של פונקציית התגמול ומודל אי הוודאות משלב 2).

איבר רגולריזציה השומר את פונקציית f הנלמדת (מה-SDE) באיטרציה הנוכחית (של האלגוריתם ולא של מודל דיפוזיה) קרובה מבחינת מרחק KL לפונקציית f מה-SDE של המודל התחלתי. בנוסף רוצים לשמור את התפלגות הדאטה באיטרציה ההתחלתית הנלמדת v קרובה להתפלגות הדאטה ההתחלתית של המודל שהתחלנו ממנו מבחינת KL. שני הקירובים הלא צריכים להתקיים מעל כל האיטרציות של מודל דיפוזיה (פתרון של ה-SDE).

אותם איברי הרגולריזציה עבור f ועבור v שלא ״מאפשר״ להם לסטות יותר מדי מה- f ומה-v מהאיטרציה הקודמת של האלגוריתם עבור כל האיטרציות של מודל דיפוזיה.

מאמר קצת מורכב מתמטית - מקווה שעזרתי לכם קצת להבין אותו.

https://arxiv.org/abs/2402.16359