המאמר היומי של מייק - 26.12.24:  
RL for Consistency Models: Faster Reward Guided Text-to-Image Generation

מזמן לא סקרתי מאמרים על מודלי דיפוזיה אז אחרי שנתקלתי במאמר הנחמד המשלב מודלי דיפוזיה גנרטיביים עם למידה עם חיזוקים (Reinforcement Learning או RL בקצרה), לא היו לי ספקות שזה הולך להיות המאמר המסוקר. כאמור המאמר פיתח שיטת אימון מודל של דיפוזיה גנרטיבי מסוג Consistency Model או CM.

קודם כל נשאלת השאלה למה צריך לאמן מודלי דיפוזיה גנרטיביים עם שיטות הלקוחות מעולם RL. הרי יש לנו שיטות סטנדרטיות יותר לאימון של מודלי דיפוזיה שהצליחו להביא לנו מודלים בעלי ביצועים מרשימים (בגנרוט תמונות מטקסט). אתם בטח יודעים שאימון מודלי דיפוזיה לגנרוט תמונות זה דבר לא זול ודורש לא מעט זמן ושימוש RL לאימון (או fine-tune) של מודלי דיפוזיה יכול לחסוך לנו זמן במקרים שאנו צריכים לאמן מודל דיפוזיה ייעודי (למשל לדומיין נישתי)

אחת הדוגמאות למשימה כזו היא אימון מודל ליצירת תמונות מפרומפט (תיאור טקסטואלי) כאשר יש בידינו פונקציה המשערכת את התאמת התמונה לפרומפט. אתם כבר יכולים לנחש שפונקציה זו תשרת לנו בתור פונקצית תגמול (reward function).

כבר הזכרתי שהמאמר משלב שיטה חדשה (יחסית) לאימון מודלי דיפוזיה הנקראת CM ושיטה זו (שהומצאה על ידי איליה סלוצקב ושות') מאפשרת גנרוט יותר מהיר של מודלי דיפוזיה גנרטיביים. בגדול מאוד שיטה זו מנסה לאמן מודל שאוכף עקביות בין התמונות המשוחזרות על ידי המודל מתמונות מורעשות עם עוצמות שונות רעש. כלומר לוקחים תמונה, מרעישים אותה עם רעש (בד״כ גאוסי) עם שונויות שונות ומאמנים מודל להחזיר את אותה התמונה הנקייה (עקביות לשמה).

למה השיטה הזו מאפשרת גנרוט יותר מהיר של תמונות? כי בגדול היא מאפשרת לגנרט תמונה נקייה מרעש באיטרציה אחת בלבד (ככה המודל מאומן). במציאות עושים את זה בכמה איטרציות (מספר קטן). מתחילים מרעש, מגנרטים את התמונה ממנו, מוסיפים פחות רעש לתמונה המגונרטת, מגנרטים מהתמונה המורעשת שוב וממשיכים ככה כמה איטרציות (עשרות בודדת). זה מאפשר לזרז את תהליך הגנרוט כי מודלי דיפוזיה סטנדרטיים צריכים מאות איטרציות בד״כ.

אוקיי, אחרי הקדמה ארוכה נעבור לתיאור של מה שעשו במאמר. המחברים הגדירו Markov Decision Process c או MDP המתאר תהליך גנרוט של תמונה (או כל דאטה אחר למעשה). כאמור פונקציה תגמול ניתנת לנו והיא מודדת מידת התאמה של התמונה המגונרטת לפרומפט. המאמר מגדיר:

המצב s\_t בתור שלישיה התמונה מגונרטת באיטרציה t, עוצמת הרעש והפרומפט c

הפעולה a\_t היא התמונה באיטרציה t + 1

הפוליסי היא זו פונקצית התפלגות מותנית של תמונה מאיטרציה t+1 בהינתן התמונה המגונרטת מאיטרציה t בתוספת רעש

המצב המתחלתי הוא רעש גאוסי סטנדרטי ופונקציית תגמול נתונה לנו

אחרי שהגדרנו את ה-MDP של תהליך גנרוט התמונה אנו יכולים להשתמש בשיטה DPO או Direct Preference Optimization לאימון פונקצית עקביות (= המודל שאנו מאמנים). למעשה DPO מאמן מודל הממקסם את פונקצית התגמול תוך כדי הגבלת של גודל עדכון פרמטרי המודל בכל איטרציה (הומצא על ג'ו שולמן ה-CTO של OpenAI לשעבר).

המאמר גם טוען שאימון כזה הוא חסכוני מבחינת משאבי החישוב הנדרשים ויעיל מבחינה הדאטה (כלומר יכול לעבוד לדאטהסטים קטנים).

https://arxiv.org/abs/2404.03673