המאמר היומי של מייק: 15.04.25  
Classifier-Free Guidance inside the Attraction Basin May Cause Memorization

חוזרים לסקור מאמרי דיפוזיה - הפעם מאמר קליל (יחסית למאמר ממוצע בנושא מודלי דיפוזיה). המאמר מציע שיטה למניעת זיכרון או memorization באנגלית על ידי מודלי דיפוזיה. ניתן לראות בזיכרון סוג של mode collapse (הזכורה לנו מתקופת הגאנים) כאשר המודל מגנרט תמונות דומות מאוד (וגם דומות לתמונות מסט האימון) לקלטים שונים (בד״כ נדגמים מהתפלגות פשוטה לדגימה כמו גאוסית סטנדרטית).

תופעה זו מתרחשת לרוב במודלי דיפוזיה מותנים כלומר כאלו שיודעים לצייר לנו תמונה מתיאור טקסטואלי (כלומר פרומפרט). במקרה זה תופעת זיכרון מתרחשת כאשר לא משנה מאיזה דגימה התחלתית של רעש גאוסי אנו מתחילים, המודיל מגנרט לנו תמונות כמעט זהות. המאמר החוקר את הסיבות להתרחשות תופעה זו ומגיע למסקנה כי זיכרון קורה עקב שימוש בטכניקה הנקראת Classifier Free Guidance או CFG בקצרה.

המטרה של CFG היא ״להזיז״ את גנרוט התמונה לכיוון הסמנטי של הפרופמט כלומר לגרום לתמונה להיות מותאמת לפרומפט. אתם בטח יודעים שמודלי דיפוזיה מגנרטים תמונה עלי ידי הסרת רעש הדרגתית מהרעש הטהור (בד״כ גאוסי). זה מתבצע כאמור באיטרציה באמצעות מודל דיפוזיה שמאומן לשערך את הרעש שצריך להסיר בהינתן תמונה מורעשת באיטרציה t (נציין כי t הוא גם קלט למודל דיפוזיה).

כאמור CFG ״מזיז״ את התמונה המגונרטת לכיוון הפרומפט על ידי הוספת הרעש המשוערך על ידי מודל את ההפרש ממושקל (עם משקל קטן) בינו (הרעש המשוערך) לבין הרעש המשוערך של מודל דיפוזיה לא מותנה (שמאומן לגנרט תמונה ללא פרומפט). גם המודל (בזמן האינפרנס) מזיז את התמונה המגונרטת רחוק יותר מהתמונה הממוצעת (ללא פרומפט) ומקרבת אותו (סמנטית) לפרומפט שלה.

אבל כמו שהמחברים מצאו CFG מקרב את התמונה לפרומפט חזק למדי. יתרה מזו הם מצאו שאם מתחילים לעשות CFG מאיטרציה מאוחרת יחסית (כאשר התמונה כבר נוקתה קצת מהרעש) אז תופעת הזיכרון כמעט ולא מתרחשת. הסיבה לכך טמונה בכך שהנורמה של וקטור הרעש המותנה גבוהה משמעותית מזו שאינה מותנית באיטרציות מוקדמות אך הן משתוות לקראת אמצע תהליך של הסרת הרעש (backward process).

אז בשביל להתמודד עם תופעת הזיכרון המאמר מציע לעשות הסרת רעש ללא CFG באיטרציות מוקדמות ולהתחיל עם CFG באיטרציות יחסית מאוחרות. אבל איך ניתן לזהות מתי צריך להתחיל להפעיל CFG? פשוט מאוד - כאשר המרחק בין נורמות הרעשים המשוערכים מתחיל לקטון. זה בגדול הרעיון העיקרי של המאמר.

במאמר יש לא מעט הגדרות מתמטיות להגדרת הזיכרון (ואני מאוד אוהב את זה) אבל מי שלא רוצה להתעמק יכול להסתפק בסקירה זו להבנה כללית.

https://arxiv.org/abs/2411.16738