המאמר היומי של מייק - 16.02.25  
Representation Alignment for Generation: Training Diffusion Transformers is Easier than you Think

לוקחים פסק זמן קטן מ-LLMs וסוקרים מאמר על מודלי דיפוזיה גנרטיביים. המאמר מציע שיטה די אינטואיטיבית לשיפור ביצועים של מודלים אלו על ידי הוספת איבר רגולריזציה ה״מיישר״ את הייצוגים הפנימיים של המודל עם אלו של אנקודרים חזקים כמו DiNOV2. יישור זה משפר את איכות התמונות שהמודל מגנרט.

נתחיל מרקע קצרצר על מודלי דיפוזיה גנרטיביים. מודלים אלו מאומנים לגנרט תמונות (למשל בהינתן תיאור טקסטואלי) על ידי הסרה הדרגתית של הרעש. המודל מתחיל מרעש טהור (בד״כ גאוסי) ולאט לאט הופכים אותו לתמונה (או פיסת דאטה מדומיין אחר). המודל מאומן על תמונות מורעשות עם רמות שונות של רעש(=איטרציות) כאשר באימון המודל לומד להסיר כמות קטנה של רעש (מאיטרציה t לאיטרציה t -1). בחירה של הייפר-הפרמטרים של תהליך ההרעשה היא מרכיב קריטי לאיכות גנרוט של המודל המאומן.

תהליך זה(הרעשה) ניתן לתאר באמצעות משוואות דיפרנציאלית של זרימה הסתברותית (probability flow) המתאר השתנות (גרדיאנט) הדאטה המורעש עם קצב/מהירות הרעשה (velocity) שנסמן אותו (הפתרון של משוואה זו מתפלג לפי ההתפלגות של הדאטה המורעש). קצב הרעשה ניתן לשערך עם המודל (=רשת) בהתבסס על דגימות הדאטה המורעש ו-. לאחר מכן ניתן לפתור את משוואות הזרימה ההסתברותית עם השערוך של (בכיוון ההפוך - כלומר החל מרעש טהור) עם שיטת איולר למשל. שיטות אלו נקראות stochastic interpoland. נציין שיש שיטות המבוססות על פתרון נומרי של משוואה דיפרנצאלית סטוכסטית שמתארת את השתנות הדאטה כפונקציה של פונקציית score שהיא לוגריתם של פונקציית התפלגות של דאטה מורעש.

אוקיי, אחרי הסיבוך הזה החיים נהיים קצת יותר קלים. מודלי דיפוזיה היום הם לרוב מודלים לטנטים כאשר הגנרוט מתרחש במרחב הייצוג של הדאטה. כלומר המודל מאומן לשחזר ייצוג לטנטי מרעש ואז מפעילים את הדקודר כדי לבנות תמונה מהייצוג המשוחזר. הייצוג של התמונה ההתחלתית נוצר על ידי האנקודר. המחברים טוענים שהייצוגים הלטנטיים המורעשים אינם ״חזקים מספיק״ כלומר פחות משקפים את האספקטים הסמנטיים של התמונה.

המחברים מציעים להעשיר את הייצוגים האלו על ידי הוספה של איבר רגולריזציה שמטרתו לקרב ייצוגים אלה (של התמונת המרועשות) לייצוג המופק על ידי אנקודר חזק (כמו DINOV2). לוס זה מתווסף ללוס הרגיל של מודל דיפוזיה ונטען במאמר שזה משפר את איכות התמונות המגונרטות וגם תורם ליציבות האימון.

https://arxiv.org/abs/2410.06940