המאמר היומי של מייק: 17.04.25  
Memorization to Generalization: The Emergence of Diffusion Models from Associative Memory

אוקיי, ממשיכים עם מאמר תאורטי עמוק בנושא מודלי דיפוזיה גנרטיביים. בסקירה הזו ניסיתי למקסם את אחוז המושגים של ML שתרגמתי לשפת הקודש. תגידו לי איך יצא.

המאמר מציג דיון תיאורטי עמוק, הרואה במודלי דיפוזיה מערכות זיכרון אסוציאטיבי סטוכסטיות בעלות מספר פרמטרים עודף (overparameterized). הרעיון המרכזי הוא שהתנהגותם של מודלי דיפוזיה כתלות בגודל דאטהסט האימון משקפת את הדינמיקה של רשתות הופפילד מודרניות(התפתחות של אלו שהוצאו על ידי חתן פרס נובל טרי) כאשר הן חורגות מקיבולת הזיכרון הקריטית שלהן.

רשת הופפילד הקלאסית היא מודל של זיכרון אסוציאטיבי שבו כל תבנית נשמרת כנקודת מינימום באנרגיה, אך הקיבולת שלה מוגבלת – היא יכולה לשמור רק מספר תבניות פרופורציונלי למספר הנוירונים בה. רשת הופפילד מודרנית מרחיבה את הרעיון באמצעות מנגנון softmax או אגרגציה לא ליניארית, ומסוגלת לשמר כמות אקספוננציאלית של תבניות ולשחזרן בדיוק גבוה, תוך קשר הדוק למכניקת מנגנון ה-attention בטרנספורמרים (סקרתי מאמר על זה).

ברשת הופפילד קלאסית, משטח האנרגטי בנוי כך שכל תבנית מאומנת מהווה נקודת משיכה יציבה. כל עוד מספר התבניות נמוך מהקיבולת התיאורטית, כל תבנית (וקטור או דגימה שצריך לזכורו) מוקצית לבור אנרגטי מבודד. כאשר מספר התבניות חורג מהקיבולת, מופיעות נקודות משיכה לא צפויות — מה שנקרא "מצבים מזויפים"(spurious). מצבים אלו אינם תואמים לדגימות האימון, אך לעיתים קרובות מהווים קומבינציות לינאריות שלהן.

המאמר מזהה תופעה דומה במודלי דיפוזיה. במהלך האימון, מודל הדיפוזיה לומד פונקציית ציון (score) עבור תהליך הופכי (backward) בתהליך הפיכת רעש טהור לפיסת דאטה. פונקציה זו מקודדת נגזרת של הסתברות לוגריתמית של פיסת דאטה מורעשת וכתוצאה מכך ניתן לראות בה את שיפוע פונקציית אנרגיה סמויה (הסתברות גבוהה מתאימה לאנרגיה קטנה - סוג של אטרקטור). המחברים מראים שפונקציה זו זהה בצורה לזו של רשת הופפילד מודרנית עם אגרגציה מבוססת softmax על הדאטה (צורה בה רשת הופפילד זוכרת את הדאטה). כלומר, הדינמיקה של דיפוזיה שקולה למינימיזציה סטוכסטית של במשטח אנרגיה של זיכרון אסוציאטיבי.

כאשר דאטהסט האימון קטן, פונקציית score מחושבת בדיוק גבוה(המודל overparameterized ומשערך אותה בקלות) ורוב הדגימות שנוצרות הן העתקות של דגימות האימון - המודל מצוי בשלב של זיכרון(memorisation) חזק. ככל שגודל הדאטה גדל, המודל כבר לא יכול לייצר בורות אנרגיה מבודדים לכל דגימה, ונוצרים מצבים "מזויפים״. אלו הם דגימות שלא נראות בסט האימון אך כן נמצאות קרוב אליהם ומהוות סוג של שילובים שלהם. בהמשך, כשקיבולת זו (של שילובים) מנוצלת גם היא, המודל מתחיל לייצר דגימות חדשות שלא שייכות לא לסט האימון ולא לקבוצת השילובים - זהו שלב הכללה מלאה.

המאמר מגדיר שלוש קיבולות:

קיבולת הזיכרון: מספר הדגימות המרבי שמודל יכול לשחזר באופן עקבי מתוך האימון.

קיבולת הזיוף - spurious (שילוב): גודל הדאטה שבו יש מקסימום שכיחות לדגימות שלא מופיעות באימון אך כן מופיעות בקבוצת הסינתזה.

קיבולת ההכללה – גודל קבוצת האימון שמעליו המודל מפסיק לייצר דגימות שכפולות או קרובות לדאטה הקיים.

המעבר בין שלבים אלה מאופיין כהתנהגות פאזה: תחילה ירידה חדה בזיכרון, עלייה חדה ב״זיופים״, ואז דומיננטיות של דגימות כלליות. המאמר מיישם מדדי זיהוי מבוססי שכנות קרובה למדידת המרחק בין הדגימות שנוצרו לבין הדאטה המקורי, ומסווג לפיו אם מדובר בזיכרון, ״שילוב״ או הכללה.

מבחינה תיאורטית, העבודה מצביעה על כך שהכללה אינה נובעת רק ממבנה הארכיטקטורה או מהרגולריזציה, אלא מתוך אינטראקציות מבניות במשטח האנרגטי. כאשר כמות הדאטה חורגת מהקיבולת, בורות האנרגיה מפסיקים להיות דיסקרטיים ומתחילים ליצור משטח רציף - שילובים הם תוצר ישיר של אינטראקציות אלו. ככל שהאינטרפולציה ביניהם(השילובים) משתפרת, נוצרות ״משטח אנרגטי מכליל״ שהוא תוצאה של דינמיקה אגרגטיבית של נקודות המשיכה מהדאטהסט.

המסקנה המרכזית היא שמודל דיפוזיה פועל בפועל כמערכת זיכרון אסוציאטיבי רוויה, והכללה נוצרת לא כתכונה חיצונית אלא כתוצר של קריסת קיבולת זיכרון — תופעה הניתנת לאפיון, כימות וחיזוי.

https://openreview.net/forum?id=zVMMaVy2BY