המאמר היומי של מייק - 16.01.25  
Diffusion Models for Non-autoregressive Text Generation: A Survey

היום נסקור סקירה מלפני שנה וחצי של תחום (משפחת טכניקות) אז מטבע הדברים זה הולך להיות די קצר. הסקירה היא על שיטות גינרוט טקסט לא אוטורגרסיביות כלומר לא טוקן אחרי טוקן אלא סדרה שלמה. השיטות שנדבר עליהן מגנרטות טקסט בכמה איטרציות אבל זה לא נעשה בצורה אוטורגרסיבית - למשל שיטות אלו יכולת לגנרט טוקן מספר 78 לפני טוקן מספר 24.

אוקיי, בטח כמה מכן חשבו על מודלי דיפוזיה גנרטיביים אחרי שהזכרתי שיטות איטרטיביות ואתם לא טועים כאן. בסקירה קצרה זו אסביר בצורה מתומצתת אין ניתן לגנרט טקסט עם מודלי דיפוזיה. כמו שאתם בטח זוכרים מודלי דיפוזיה מאומנים להסיר רעש מדאטה מורעש וזה נעשה באיטרציות. כלומר המודל מאומן להסיר כמות קטנה של רעש מהדאטה עד להגעה לדאטה נקי וכך לאחר האימון המודל מסוגל לגנרט דאטה מרעש טהור בכמה איטרציות.

אבל איך ניתן להוסיף רעש לטקסט שחי במרחב דיסקרטי (כלומר טוקנים). יש בגדול שתי גישות: הגישה הרציפה והגישה הדיסקרטית. בגישה הרציפה שהיא יותר פשוטה וקרובה ליבנו אנו לא פועלים במרחב הדיסקרטי אלא במרחב של אמבדינגס. בגישה הרציפה אנו הופכים את הטקסט שלנו לוקטור אמבדינג רציף אבל להבדיל אנקודר רגיל אנו הופכים כל טוקן לייצוגו הווקטורי בנפרד מהאחרים. לאחר מכן מאמנים מודל דיפוזיה לגנרט אמבדינג של טקסטים. הוספת רעש ואימון מודל denoising מתרחשים במרחב האמבדינג כאשר המטרה היא הסופית היא לשחזר את הטוקנים מהאמבדינגס (ד״א יש כמה שיטות לעשות את זה) אחרי ניקוי רעש.

משפחת השיטות השנייה היא לבצע הוספת רעש במרחב הדיסקרטי. מובן שהרעש לא יכול להיות רציף אז מה שניתן לעשות היא לשנות את ערכי הטוקנים (למשל לטוקן [mask]) בהסתברות מסוימת כאשר המטרה היא באיטרציה האחרונה להפוך את כל הטוקנים ל-[mask]. מודל דיפוזיה באיטרציה i מאומן לחזות את הטוקנים מהאיטרציה הקודמת, כאשר באינפרנס הגנרוט מתחיל מכך שכל הטוקנים שווים ל-[mask] והמודל לאט לאט הופך אותם לטקסט.

כמובן שאופן הרעשה של טוקן בכל איטרציה זה הייפרפרמטר השקול ל-noise schedule במודלי דיפוזיה רגילים. ניתן לתאר אופן הרעשה בתור מטריצה. כל טוקן ניתן לייצוג על ידי וקטור ההסתברות (מעל מילון הטוקנים) אז ניתן לייצוג טוקן מאיטרציה i כמכפלה פנימית של ייצוגו באיטרציה i-1 על ידי מטריצה סטוכסטית Q\_i (סכום של שורות ועמודות הינו 1). Q\_i היא הייפרפרמטר הכי חשוב במודלי דיפוזיה דיסקרטיים.

מתברר שזה תחום מחקר די פעיל למרות עדיין מודלים אלו לא הגיעו לביצועים של מודלי שפה אוטורגרסיביים. אבל אני לא פוסל שזה עוד יקרה כי מודלים אלו מסוגל לעבוד בתפוקה גבוהה יותר ממודלים אוטורגרסיביים (עבור מספר צנוע של איטרציות).

https://arxiv.org/abs/2303.06574