# Review 30: Diffusion-LM Improves Controllable Text Generation

**Paper: https://arxiv.org/abs/2205.14217v1**

פינת הסוקר:

המלצת קריאה מכותבי הסקירה: מאמר מומלץ למתעניינים במודלי שפה, במיוחד לאלה שמתמקדים בגנרוט טקסט מוכוון משימה (controllable text generation)

בהירות כתיבה: בינונית

ידע מוקדם:

יסודות מתמטיים של מודלי דיפוזיה הסתברותיים לגנרוט דאטה (Denoising Diffusion Probabilistic Models -DDPM)

הבנה טובה במודלי שפה ליצירת טקסט מוכוון משימה

יישומים פרקטיים אפשריים:

יצירת טקסט מוכוון משימה איכותי יותר

פרטי מאמר:

פורסם בתאריך: 27.05.22, בארקיב.

הוצג בכנס: טרם ידוע

תחומי מאמר:

מודלים גנרטיביים ליצירת טקסט מוכוון משימה

מודלי דיפוזיה לגנרוט פיסות דאטה חדשות

כלים מתמטיים, מושגים וסימונים:

Denoising Diffusion Probabilistic Models -DDPM

מודלי שפה מבוססי רשתות נוירונים

תמצית מאמר:

מבוא:

מודל דיפוזיה הסתברותי (DDPM - Denoising Diffusion Probabilistic Model) שייך למשפחת המודלים הגנרטיביים המאפשרים יצירת פיסות דאטה חדשות. משפחה זו כוללת גם (Variational AutoEncoders (VAE, גאנים (GAN - Generative Adversarial Networks), מודלים של זרימה מנורמלת (Normalized Flows) וכמה מודלים גנרטיביים ״פופולריים״ קצת פחות. בשנה האחרונה הצליחו מודלי דיפוזיה להשוות ואפילו לשפר את ביצועי מודלי SOTA בתחום הראיה הממוחשבת (בעיקר גאנים ו- VAE-ים). מודלים אלה שימשו ליצירה של תמונות באיכות מרהיבה (ראה למשל Diffusion Models Beat GANs on Image Synthesis שנסקר במדור זה). בנוסף לפני כמה חודשים עלה לאוויר מאמר שהציע שיטה בשם GLIDE שהצליח ליצור תמונות מדהימות מתיאורם באמצעות מודל דיפוזיה. לאחר מכן יצאו כמה עבודות שהשתמשו במודלי דיפוזיה במשימות שונות של הראיה הממוחשבת למשל עבור super-resolution (שיפור הרזולוציה של תמונה).

עכשיו נשאלת השאלה האם ניתן לנצל פרדיגמה זו למשימות של עיבוד שפה המערבות גנרוט טקסט למשל ליצירת טקסט מכוון משימה? המאמר הנסקר מספק תשובה חיובית לשאלה זו ומציע שיטה לגנרוט טקסט מכוון משימה באמצעות מודיפיקציה של מודל דיפוזיה.

כעת נתאר בקצרה את יסודות מודלי הדיפוזיה ההסתברותיים לגנרוט דאטה (ראה סקירה זו לדיון יותר מעמיק). מודל דיפוזיה מורכב משני ״תהליכים״: התהליך הקדמי והתהליך האחורי. התהליך הקדמי מורכב מכמה איטרציות של הוספת רעש גאוסי לדאטה כאשר מטרת כל איטרציה היא לטשטש את הדאטה. למשל אם הדאטה היא תמונה אז כל פיקסל עובר תהליך דעיכה: התוחלת של של ערך הפיקסל לאחר איטרציה t הופכת להיות כאשר והשונות שלו גדלה. מטרת התהליך הקדמי היא ״להפוך״ פיסת דאטה לרעש גאוסי איזוטרופי כלומר מפולג (N(0, I. המטרה של התהליך האחורי היא לשחזר את הדאטה מהרעש - וזה גם נעשה באיטרציות. מודל המסוגל לבנות פיסות דאטה מרעש (בתהליך איטרטיבי) הוא למעשה מודל גנרטיבי לכל דבר.

אבל איך זה נעשה? קודם כל בדומה למה שמקובל באימון מודלים גנרטיביים נרצה ״להתאים״ מודל פרמטרי הממקסם את פונקציית הנראות המירבית (likelihood) של הדאטהסט. ניתן להשתמש ב- ELBO בשביל לבנות את החסם התחתון לנראות המירבית:

כאשר (q(xt-1|xt, x0 היא ההתפלגות המושרית באמצעות התהליך הקדמי (והיא גאוסית!!), היא המודל הפרמטרי שמאומן לשערוך xt-1 מ-xt ו- x0 היא הדוגמא המקורית). כלומר מטרת האיטרציה t של התהליך האחורי היא לחזות את ערך הדאטה באיטרציה הקודמת ,t-1, כלומר אנו רוצים לשערך את (q(xt|xt-1. כאן xt מסמן פלט של האיטרציה t של התהליך הקדמי.

הערה חשובה: חשוב לציין כי למרות שהתפלגות (q(xt|xt-1 היא גאוסית לפי הגדרת התהליך הקדמי (q(xt-1|xt אינו גאוסי. (q(xt-1|xt תלויה גם בהתפלגות של הדאטה עצמו - ניתן לראות את זה בקלות אם פותחים (q(xt-1|xt באמצעות נוסחת בייס. אולם עבור בחירה מושכלת של קבועי ניתן לקרב (q(xt-1|xt באמצעות התפלגות גאוסית וזה בדיוק מה שנעשה במודלי דיפוזיה.

מודלי דיפוזיה גנרטיביים הראשונים השתמשו בפונקציית הלוס הנ״ל אך לא הצליחו להציג ביצועים ברי השוואה עם שיטות SOTA. ועקב כך במאמרים מאוחרים יותר במקום למזער KL-divergence מאמנים מודל לחזות את התוחלת המותנית של הדאטה באיטרציה 1-t. תוחלת זו מסומנת ב-, כלומר המודל מאומן למזער את הביטוי הבא:

הערה: יש כמה עבודות שבמקום לחזות את התוחלת של הדאטה בשלב הקודם , בנו מודל לחיזוי של הרעש שהתווסף בשלב t (רפרמטריזציה בעצם). גישה זו הוכחה כיציבה יותר מחיזוי ישיר של התוחלת (ראיתי כמה הסברים לגבי הסיבה לתופעה זו אך לא השתכנעתי).

הסבר של רעיונות בסיסיים:

מודל דיפוזיה מותאם למודל שפה (לא מכוון משימה):

אז איך נוכל לבנות מודל שפה מבוסס על עקרון הדיפוזיה? הרי הקלט למודל הזה הוא דיסקרטי (המורכב ממילים או טוקנים) והוספת רעש במרחב הזה לא ניתנת להגדרה פשוטה וגם משמעותה לא לגמרי ברורה. במקום זאת המחברים מציעים להשתמש בשיכונים (embeddings) ולמרות שמרחב השיכונים הוא גם דיסקרטי בטבעו, הוספת רעש למרחב שיכון היא יותר טבעית ו-tractable. למעשה בשביל להתאים את מודל הדיפוזיה למשימה הזו, המחברים הוסיפו שלב ראשוני של חישוב שיכון טוקנים.

כאמור, בתהליך הקדמי קודם כל ממפים את המילים w לוקטורי השיכון x0 כך שהתפלגות של x0 מוגדרת בתור (qφ(x0|w)~ N(Emb(w), σ0I. לתהליך האחורי (denoising) מוסיפים שלב נוסף בסוף הנקרא עיגול (rounding). מטרתו של שלב העיגול היא למפות את וקטורי השיכון x0 מהשלב האחרון של התהליך האחורי לטוקנים מהמילון (למעשה לטקסט). פעולה זו נעשתה באמצעות שערוך של (pθ(w|x0.

המאמר מראה כי פונקציית הלוס של מודל הדיפוזיה (החוזה תוחלת של האיטרציה הקודמת) עם השלבים שנוספו תיראה באופן הבא:

למעשה האיבר השני בנוסחה הזו ממזער את ההפרש בין השערוך של השיכון של הדאטה המקורי x0 מהדאטה של האיטרציה הראשונה x1 לבין השיכון של x0, המסומן (EMB(w. האיבר האחרון ממקסם את ההסתברות של הטוקן w בהינתן השיכון שלו x0.

מהלך אימון של מודל שפה דיפוזיוני:

אימון מודל השפה הדיפוזיוני עם פונקצית הלוס הזו, טומן בחובו כמה מוקשים (איכות הטקסט שנוצר לא גבוהה מספיק) שדורשים שימוש במספר טריקים:

טריק 1: במקום לחזות את התוחלת של האיטרציה הקודמת, המחברים הציעו לחזות את הדאטה עצמו (השיכון שלו) x0

טריק 2 (Clamping): במקום להשתמש באומדן של x0 באיטרציה t (המסומן ב-(fθ(xt, t) כדי לדגום את xt-1 ממנו, המחברים דוגמים את xt-1 מהשיכון של סדרת המילים הקרובה ביותר ל-(fθ(xt, t.

הערה: ניתן לקבל ביטוי סגור עבור xt-1 מ- x0 ורעש גאוסי סטנדרטי (N(0,1 דרך שימוש בהגדרת התהליך הקדמי.

מהלך אימון של מודל שפה דיפוזיוני מכוון משימה:

אז יש לנו כרגע מודל דיפוזיה מאומן שיודע לגנרט טקסט מרעש. השאלה עכשיו איך נוכל למנף אותו ליצירת טקסט מכוון משימה, כלומר בהינתן משתנה בקרה (control variable), המסומן c .c יכול להיות עץ סינטקטי או סנטימנט שהטקסט הנוצר צריך לקיים. קודם אנו צריכים לגנרט פלטים של מודל הדיפוזיה באופן מכוון משימה, המוגדרת על ידי c:

אז לפי חוק בייס (p(xt-1|xt, c) ∝ p(xt-1|xt) p(c|xt-1, xt. כדי ליצור פלט של שלב t-1 של מודל הדיפוזיה מבצעים gradient descent עבור ((log(p(xt-1|xt, c כאשר הפרמטר המאופטם הוא xt-1:

המאמר מציע שני חידושים קלים לתהליך אופטימיזציה זה. הראשון הוא fluency regularization המכניס משקול בין שני איברי הנוסחה: (λlog p(xt−1|xt) + log p(c| xt−1. מקדם זה תורם לויסות האיזון בין ״שטף של הטקסט״ להתאמה למשימה. החידוש השני הוא שימוש ב 3 איטרציות של GD לכל שלב של דיפוזיה.

מכיוון שתהליך הגנרוט המתואר די איטי, המחברים השתמשו ב-200 איטרציות של מודל דיפוזיה במקום 2000 המקובלות. גישה זו עדיין איטית באופן משמעותי משיטות פשוטות יותר, כמו אימון טרנספורמר במבנה encoder-decoder.

ניסויים ותוצאות

בחינת ביצועי המודל נעשתה עבור מס׳ משימות שכל אחת מהן בחנה היבט אחר של גנרוט הטקסט, כאשר המדדים להשוואה היו רהיטות (fluency) הטקסט ועמידה במשימות אליהן הוא הוכוון:

גנרוט טקסט מוכוון משימה - יצירת טקסט בהתאם לפרמטרים נתונים. למשל, טקסט עם סנטימנט חיובי.

גנרוט טקסט מוכוון משימות מרובות - יצירת טקסט בהתאם למספר פרמטרים רצויים. לדוגמה, טקסט חיובי בעל מבנה סמנטי מוגדר.

השלמה של טקסט חסר (infilling) - משימת השלמה של מקטעי טקסט חסרים אשר תואמים טקסט נתון.

גנרוט טקסט מכוון משימה

על מנת לבדוק את פלט המודל, המחברים השוו את הטקסט שגונרט לפלטים של מודלי שפה אחרים בחמש משימות שונות. בכל המשימות בהן מודל הדיפוזיה נבחן, הוא הציג תוצאות טובות יותר ממודלי שפה גדולים אחרים (FUDGE ו- PPLM). גם בהשוואה למודל (GPT-2) שאומן במיוחד למשימות אילו, להפתעת החוקרים, עלו ביצועי המודל הדיפוזי על מודל (GPT-2) שאומן במיוחד למשימות אילו בשתי משימות שדרשו הסתכלות על כל הטקסט במקביל (גנרוט טקסט בהנתן עץ תחבירי וגנרוט טקסט בהינתן טווח תחבירי נתון), כאשר ביתר המשימות המודל הדיפוזי השתווה למודל המאומן.

גנרוט טקסט עבור משימות מרובות

אספקט נוסף של גנרוט שפה מוכוון משימה הוא האפשרות לכוון את הטקסט המגונרט למס׳ משימות במקביל. החוקרים בחנו זאת ע״י השוואה לשני מודלי שפה אחרים (FUDGE ומודל PoE) בשני ניסויים שבכל אחד מהם הטקסט כוון לשתי משימות שונות. בשני הניסויים מודל הדיפוזיה הראה יכולת טובה יותר בגנרוט טקסט שעומד בקריטריונים של המשימות אליהן הוא הוכוון, אך זה בא על חשבון רהיטות הטקסט שנוצר, שם ביצועי מודל הדיפוזיה נפגעו משמעותית.

השלמת טקסט חסר

במשימת השלמת טקסט, מודל הדיפוזיה הראה ביצועים טובים יותר מכל אחד מארבעת המודלים אליהן הוא הושווה, כולל מודל שאומן במיוחד למטרה זאת. העליונות של מודל הדיפוזיה הייתה עקבית עבור כל אחת מחמשת שיטות המדידה בהן המודל נבחן.

סיכום:

המאמר מציע גישה חדשה לגנרוט טקסט מוכוון משימה באמצעות מודל דיפוזיה. גישה זו מצליחה להשיג ביצועים טובים במגוון משימות, כאשר היא מראה עליונות על גישות קיימות בגנרוט טקסט למשימות מרובות ובמשימות שדורשות הסתכלות על כל הטקסט במקביל. עם זאת, אני חושש שמהירות הגנרוט אינה גבוהה מספיק כדי להתחרות בגישות הקיימות כרגע.

שיתוף פעולה: הפוסט נכתב על ידי מיכאל (מייק) ארליכסון, PhD, Michael Erlihson, משה משען Moshe Mishan