# Review 29: Language Modeling via Stochastic Processes

**Paper: https://arxiv.org/abs/2203.11370v2**

פינת הסוקר:

המלצת קריאה מכותבי הסקירה: שווה קריאה למי שעוסק בגנרוט טקסטים באמצעות מודלי שפה

בהירות כתיבה: בינונית

ידע מוקדם:

הבנת העקרונות הבסיסיים של מודלי שפה (במיוחד במשימת גנרוט טקסטים)

רקע בתהליכים סטוכסטיים (כמו גשר בראוני – brownian bridge)

יישומים פרקטיים אפשריים:

גנרוט טקסטים ארוכים בהינתן הקשר

פרטי מאמר:

פורסם בתאריך: 21.03.22, בארקיב.

הוצג בכנס: ICLR 2022

תחומי מאמר:

מודלי שפה לגנרוט טקסטים

תהליכים אקראיים

כלים מתמטיים, מושגים וסימונים:

גשר בראוני (brownian bridge)

מבוא:

מודלי שפה מודרניים מפגינים ביצועים מצוינים בתחום גנרוט טקסטים קצרים (גם תלויי הקשר). לעומת זאת מודלים אלו מתקשים בגנרוט טקסטים ארוכים שבמקרים רבים מתקשים לשמור על קוהרנטיות וגם עלולים ״לסטות״ מהנושא (ההקשר). אחת הסיבות ליכולת ירודה זו הינה האופן האוטורגרסיבי שבו מודלי שפה מגנרטים טקסט. עם גישה זו כל טוקן (מילה או רצף מילים סמוכות) נוצר בהינתן הטוקנים שכבר נוצרו על ידי המודל. מכיוון שה-״זכרון״ האפקטיבי של המודל מוגבל ומסוגל ״לזכור״ רק כמות מוגבלת של מידע, המודל מתקשה להתחשב במשפטים הנמצאים רחוק מהמשפט שכרגע מג'ונרט. תופעה זו עשויה להוות הסיבה העיקרית לאי-קוהרנטיות של טקסטים ארוכים הנוצרים באמצעות מודלי שפה.

תמצית מאמר:

אחת הדרכים להתמודד עם הבעיה הזו היא ללמוד את ההתפלגות של הייצוג הלטנטי של מקטעי טקסט ארוכים (כגון משפט). לאחר מכן, ניתן לנצל התפלגות זאת כדי להתנות את הפלט במידע נוסף, ולהנחות את יצירת הטוקנים במהלך גנרוט הפלט. כלומר, קודם כל אנו מגרילים את השיכונים של כל משפט בטקסט ואחר כך מתנים את יצירת הטוקנים בייצוגים אלו.

ניתן להסתכל על התהליך הזה כתכנון מסלול של טיול ארוך: קודם מחליטים על אבני דרך חשובות ואז בונים את המסלול ביניהם. גישה זו מחזקת את הקוהרנטיות של הטקסט המגונרט מאחר ותכנון כזה מאפשר לשמור על הקשרים הסמנטיים בין חלקים רחוקים בטקסט וכך הסבירות של גנרוט טקסט לא קוהרנטי יורדת.

הסבר על הרעיון העיקרי:

אבל איך נוכל לבנות ״מסלול״ כזה לטקסטים ארוכים? עבודות קודמות הציעו גישות המבוססות על הלמידה הניגודית (contrastive learning) אך לטענת המאמר גישה זו מצליחה לעקוב רק אחרי קשרים בין משפטים סמוכים ומתקשה להתמודד עם מידול יחסים בין משפטים רחוקים בטקסט. במטרה לתת מענה לסוגייה זו, המחברים מציעים גישה ״גלובלית״ המנסה למדל את ההתפלגות של כל המשפטים בטקסט יחד במרחב הלטנטי. התפלגות המשפטים ממודלים באמצעות גשר בראוני (brownian bridge). גשר בראוני zt הוא מקרה פרטי של תנועה בראונית (brownian motion או standard Wiener process) מוגדר במקטע כאשר ערכי התהליך ב-t=0 וב-t=T הם דטרמיניסטיים (המסומנים z0 ו- zT בהתאמה).

הדינמיקה של גשר בראוני כזה מוגדרת באופן הבא:

כאשר N היא ההתפלגות הגאוסית. המשוואה הזו היא די אינטואיטיבית – היא למעשה מבצעת סוג של אינטרפולציה לינארית רועשת בין z0 ו-zT . למשל ככל ש-t יתקרב ל-T, התוחלת של zt תתקרב ל-zT והשונות תקטן. נעיר כי השונות המקסימלית מתקבלת בנקודה t=T/2 שבה באופן טבעי יש מקסימום אי-וודאות. השיטה המוצעת נקראת Time Control – TC.

המטרה כאן למצוא פונקציה

הממפה משפט למרחב הלטנטי כך שההתפלגות במרחב הלטנטי תקיים את הדינמיקה של הגשר הבראוני. כמובן, ממודלת באמצעות רשת נוירונים. כדי לאמן את הרשת הזו משתמשים בלמידה הניגודית באופן הבא:

דוגמים מיני-באצ'ים המורכבים משלשות של משפטים () כאשר המשפט חייב להיות בין ו- בטקסט. משפטים אלו יהיו הדוגמאות החיוביות.

דוגמים שלשות של משפטים באקראי כדי ליצור דוגמאות שליליות.

מבצעים אופטימיזציה עם פונקצית המחיר הניגודית הבאה:

האיבר השני במשוואה התחתונה הוא התוחלת מהביטוי עבור הגשר הבראוני.

המטרה בפורמולציה הזאת היא לכפות על ייצוגים של משפטים מאותו הטקסט לקיים את הדינמיקה של הגשר הבראוני. כלומר ייצוגים אלו צריכים להוות מסלול חלק בין ההתחלה לסוף של הטקסט במרחב הלטנטי. אלו למעשה אבני הדרך שלנו שעליהם נבסס את גנרוט הטוקנים הסופי.

אימון מודל לגנרוט טקסט עם TC:

אחרי שהבנו איך בונים את מסלול המשפטים באמצעות TC, נתאר כעת איך מגנרטים בהתבסס על מסלול זה. לטקסט נתון בעל T משפטים () ו-W טוקנים, קודם כל בונים את המסלול באמצעות אנקודר מאומן . המחברים בחרו לנצל את המסלול שנבנה כדי לכייל את GPT-2 (כלומר את הדקודר של GPT-2) שהוא מודל אוטורגרסיבי שאומן על כמות דאטה עצומה. כדי לגנרט טוקן t השייך למשפט, משתמשים בייצוג הלטנטי של המשפט ובטוקנים שגונרטו לפני כן . תהליך גנרוט הטקסט מתואר בציור למטה.

הערה לגבי כיול של GPT-2: כדי לכייל מודל שפה גדול כמו GPT-2 למשימה נתונה עם דאטהסט D לוקחים GPT-2 מאומן ומאמנים אותו על D כאשר רק חלק מהמשקלים של GPT מעודכנים והשאר מוקפאים.

גנרוט טקסט עם TC:

גנרוט טקסט עם TC הוא מתבצע באופן די טבעי. אם נתונים לנו משפט ההתחלה והמשפט האחרון אז:

מחשבים את השיכונים של ו- של המשפט הראשון והאחרון.

דוגמים את מסלול המשפטים () לפי ההתפלגות של הגשר הבראוני.

מייצרים טקסט בהינתן מסלול זה באמצעות הדקודר המכויל של GPT-2.

אם המשפט הראשון והאחרון לא ידועים, דוגמים את ייצוגיהם הלטנטיים מההתפלגויות שלהם המשוערכות על סט האימון. במילים אחרות, הפרמטרים של שיכון המשפט הראשון משוערכים על בסיס השיכונים של כל המשפטים הראשוניים של הטקסטים מהדאטהסט.

הישגי המאמר וניתוח תוצאות:

איכות הטקסט שגונרט נבחנה על ידי בחינת הדינמיקה והרהיטות של הטקסט המגונרט, עבור מקטעי טקסט לוקאליים וגלובליים.

דינמיקת טקסט מקומי

דינמיקת טקסט מקומי נבדקה על ידי בחינת היכולת של מסווג לנבא מסגרת קוהרנטית לשיח, כאשר קוהרנטיות של שיח יכולה להימדד על ידי יכולת של מודל לזהות האם זוג משפטים מופיעים בסדר תקני או לא.

בחינת היכולת של המודל לזהות האם זוג משפטים מופיעים בסדר תקין נעשתה על ידי לקיחה של שני משפטים ו כאשר {1,5,10}K וחישוב השיכונים שלהם , . אילו הוכנסו בסדר אקראי למסווג לינארי. תוצאות הדיוק (accuracy) של המסווג הליניארי הושוו לתוצאות סיווג ליניארי מחמישה מודלים שונים (GPT2, BERT, ALBERT, S-BERT, Sim-CSE) בשלושה דאטהסטים שונים (Wikisection, TM-2, TicketTalk). בכל הניסויים מודל TC הראה יכולת טובה לזהות את הסדר התקין של המשפטים, כאשר כמעט בכל המקרים דיוק המודל היה גבוה מהמודלים אליהם הוא הושווה.

רהיטות של טקסט מקומי

קוהרנטיות של טקסט מקומי נבחנה על ידי בחינה של ביצועי המודל במשימת השלמת טקסט חסר (infilling).

הדאטהסט שנבחר למשימה זאת, ROCStories, מורכב מסיפורים קצרים בני חמישה משפטים. מודל TC יצר את הגשר הבראוני על ידי חישוב השיכונים של המשפט הראשון והאחרון של כל סיפור, ובהתאם למסלול המשפטים שחושב המודל, ייצר טקסט. קוהרנטיות הטקסט שנוצר נבחנה בארבע שיטות שונות (BLEU, ROUGE, BLUERT, BERTScore) והושווה לשני מודלים שונים (LM, ILM). מודל TC קיבל ציון BLEU גבוה יותר בהשוואה למודלים אחרים, וציון דומה או מעט יותר גבוה בשאר המבחנים. בנוסף למבחנים האילו קוהרנטיות הטקסט נבחנה גם על ידי הערכה אנושית, שם המודל קיבלת ציון מעט יותר נמוך מ ILM.

דינמיקה של טקסט גלובלי

דינמיקה של טקסט גלובלי נבחנה על ידי בחינת היכולת של המודל לחקות מבנה של מסמך. לצורך זה החוקרים השתמשו בדאטהסט Wikisection, דאטהסט המכיל מסמכים ארוכים בעלי מבנה אחיד ומוגדר היטב, כאשר החוקרים בדקו האם הטקסט שגונרט על ידי המודל בכל אחד מפרקי המסמך תואם באורכו לאורך הטקסט בדאטהסט.

הטקסט שגונרט על ידי מודל TC הושווה לטקסט ששגונרט משלושה מודלים אחרים, כאשר ההשוואה היתה בכמה שונה, באחוזים, אורך הטקסט שגונרט מכל אחד מהמודלים מאורך הטקסט המקורי בכל אחד מפרקי המסמך. מודל TC הראה את חוסר התאימות הנמוך ביותר ביחס לאורך הטקסט בכל פרק ביחס לכל מודל אליו הוא הושווה.

רהיטות של טקסט גלובלי

הערכת רהיטות של טקסט גלובלי נעשה על ידי בחינה של היכולת של המודל לשמור על מבנה ברור של הטקסט ביחס לדאטהסט עליו אומן המודל ועקביות של אורך הטקסט המגונרט (לדוגמא, בדאטהסט שמכיל שיחה בין סוכן ללקוח, הצפייה היא שלאורך זמן המודל ישמור על מודל השיחה בין הסוכן ללקוח ולא יהפך למונולוג ארוך של אחד הצדדים), בכל אחד מהמבחנים האילו מודל TC הצליח להשיג ציונים גבוהים, כאשר ברובם הוא מקבל ציון גבוה ממודלים אחרים.

בנוסף לזה נעשתה גם הערכה אנושית לאיכות הטקסט המגונרט, כשמודל TC קיבלת ציון גבוה יותר ממודל GPT-2 אליו הוא הושווה.

סיכום של כל המבחנים שבדקו את הטקסט המגונרט, הראה שבעזרת מודל TC ניתן לגנרט טקסט ארוך שיהיה גם קוהרנטי וגם ישמור על מסגרת של השיח לאורך כל הטקסט המגונרט.

סיכום

מאמר זה מציע דרך חדשה לגינרוט שפה. בשונה מהגישות הקיימות כיום המתחשבות רק במצב ההתחלתי של השיחה ובאופן אוטורגרסיבי מגנרטות טקסט, מודל TC מתחשב גם במטרה הסופית של השיחה ובעזרת גשר בראוני הוא יוצר מסגרת לתוכן השיחה, בתוך מסגרת זאת הוא מגנרט את הטקסט.

תוצאות הניסויים שבוצעו לבחינת הטקסט המגונרט ממודל זה מראות שמודל TC יכול לגנרט טקסטים ארוכים תוך שמירה על קוהרנטיות ומבניות לאורך זמן.

שיתוף פעולה: הפוסט נכתב על ידי מיכאל (מייק) ארליכסון, PhD, Michael Erlihson ומשה משען.

שנקרא: