# Review 16: COCO-LM: Correcting and Contrasting Text Sequences for Language Model Pretraining

**Paper: https://arxiv.org/abs/2212.10341v2**

פינת הסוקר:

המלצת קריאה ממייק, משה: מומלץ לקרוא לאלו שעוסקים באימון מודלי שפה וגם אוהבים את הגישה הניגודית

בהירות כתיבה: גבוהה מינוס

ידע מוקדם: ידע בשיטות אימון של מודלי שפה גדולים כמו BERT וגם ביסודות הלמידה הניגודית

פרטי מאמר:

פורסם בתאריך: 27.10.21, בארקיב.

הוצג בכנס: NeurIPS.

תחומי מאמר:

עיבוד שפה טבעית (NLP)

אימון של מודלי שפה

כלים מתמטיים, מושגים וסימונים:

טרנספורמרים

Pretraining של מודלי שפה גדולים

לוס ניגודי ולמידה ניגודית

BERT, ELECTRA

מבוא:

מודלי שפה ענקיים, בעלי מיליארדי פרמטרים, המבוססים על רשתות נוירונים, הפכו להיות ברירת מחדל למרבית משימות ה-NLP בשנים האחרונות. באופן יותר מדוייק, משפחת הטרנספורמרים שולטת כמעט ללא עוררין בתחום זה. עקב הגודל והמורכבות של המודלים, קשה לאמן אותם מאפס, ומקובל לכייל מודל שאומן מוקדם יותר על דאטהסט גדול ומגוון (ובדר״כ לא מתויג), באמצעות הדאטה של המשימה החדשה. זאת, מתוך ההבנה שהשכבות המוקדמות של כל מודל נוטות להתכנס למשקלים דומים בכל המשימות באותו התחום, ולכן ניתן לחסוך במידה רבה את האימון שלהן. תהליך של אימון מודל על דאטהסט גדול נקרא בספרות אימון מוקדם (pretraining), או שהוא נקרא פשוט אימון, ותהליך הכיול נקרא אימון מחדש (retraining) או fine-tuning.

את השימוש בטרנספורמרים למשימות מידול שפה הציג לראשונה BERT לפני כ-4 שנים. BERT הציע שתי שיטות ב״פיקוח עצמי״ (self-supervised) לאימון:

מיסוך של כמה טוקנים בטקסט (בערך 15%) וחיזוים באמצעות המודל (Masked Language Modeling - MLM).

עבור שני משפטים עוקבים נתונים, המודל התבקש לחזות את סדר המשפטים.

אחד החסרונות של MLM הוא צורך בכמות גדולה מאוד של דאטה, בעיקר כי הוא מאומן על חיזוי של טוקנים ממוסכים. עבודה מאוחרת יותר הציעה שיטה בשם ELECTRA שמנסה להתגבר על קושי זה. ב-ELECTRA, במקום לחזות טוקן המודל מנסה לזהות האם הטוקן הוא ground-truth או טוקן אחר הנדגם ממודל שפה. לצורך זה מאמנים שני מודלים:

מודל שפה אוטורגרסיבי קטן יחסית שבונה התפלגות של הטוקן הבא, בהינתן הטוקנים הקודמים. מסווג זה נקרא הטרנספורמר המשני (auxiliary transformer) ומהווה MLM. המודל שהוצע בעבודה המקורית של BERT, הוא בעל פונקציית לוס softmax.

מסווג בינארי שתפקידו להבדיל בין הטוקן האמיתי לבין טוקן שונה הנדגם מהטרנספורמר המשני. השכבה האחרונה כאן היא סיגמואיד. מודל זה נקרא הטרנספורמר העיקרי (main transformer).

הערה: בשיטת ELECTRA, רק טוקנים מומסכים נדגמים מההתפלגות של הטרנספורמר המשני.

שני מודלים אלו מאומנים יחד. כך ניתן לאמן מודל שפה על כל הטוקנים ולא רק על הממוסכים ומתקבל אימון יעיל יותר הדורש פחות דאטה. המחברים מראים (בהמשך) ביצועים טובים על מספר משימות downstream. עם זאת ל-ELECTRA יש מספר חסרונות:

היעדר יכולת מובנית לגנרוט שפה. הרי המשימה של הטרנספורמר העיקרי היא דיסקרימינטיבית (סיווג) ולא גנרטיבית. זאת אומרת כדי להפוך את הטרנספורמר העיקרי למודל שפה צריך להוסיף לו מנגנון מידול שפה נפרד ולאמן אותו. המאמר טוען שמשימה זו אינה פשוטה כי האופי הדיסקרימינטיבי מגביל את יכולתו של הטרנספורמר העיקרי ללמוד פיצ'רים הנחוצים לגנרוט שפה.

ייצוגי המשפטים המתקבלים עם ELECTRA מרוכזים באזור צר של המרחב הלטנטי שלהם וייצוגי משפטים שאינם קשורים סמנטית, בהרבה מקרים, קרובים יותר אחד לשני (במובן של cosine similarity) מאשר ייצוגים של משפטים בעלי משמעות דומה.

הרעיון הכללי מאחורי המאמר:

המאמר מציע גישה הנקראת COCO-LM, המשלבת יתרונות של

MLM

הגישה הדיסקרימינטיבית של ELECTRA

גישה ניגודית (contrastive) לשיפור ייצוג של טקסט

הארכיטקטורה של COCO-LM דומה לאחת הוריאציות של ELECTRA, שנקראת All-Token MLM. שיטה זו משלבת את MLM עם הגישה הדיסקרימינטיבית ע״י שימוש בשני ראשים: הראשון דיסקרימינטיבי D, מיועד לזהות האם הטוקן, שנדגם מההתפלגות של הטרנספורמר המשני, הוא אמיתי (עם סיגמואיד כשכבה אחרונה), ואילו השני G הוא מודל שפה שמחשב הסתברות לכל טוקן בהינתן הטוקנים הקודמים (עם softmax על מרחב כל הטוקנים בשכבה האחרונה). פונקציית הלוס של ELECTRA שנבנית באופן הבא:

עבור הטוקן האמיתי ההסתברות הוא סכום משוקלל של הסתברויות של המחושבות על ידי שני ראשים: כאשר הוא הפלט של הראש הדיסקרימינטיבי (סיגמואיד).

עבור כל טוקן אחר, ההסתברות מחושבת לפי .

פונקציית הלוס במקרה הזה מוגדרת באופן הבא (לקח לי זמן להבין אותה ועקב כך הוספתי את ההסבר בפסקה הקודמת):

כאן היא הייצוג של ההקשר (הטוקנים הקודמים) הנוצר על ידי הראש הגנרטיבי G ו- הוא וקטור נלמד. מסמן טוקן שנדגם מההתפלגות המתקבלת מהטרנספורמר המשני ו- היא ״ההחלטה״ של הדיסקרימינטור (1- טוקן אמיתי ו- 0 אחרת).

תקציר המאמר:

הפרדה של למידת המשימות:

למעשה פונקציית הלוס של ELECTRA מערבבת בין שתי המטלות: הדיסקרימינטיבית (הבחנה בין טוקן אמיתי לנדגם) והגנרטיבית (בניית מודל שפה). מחברי COCO-LM טוענים שמבנה כזה של לוס מקשה על למידה בו זמנית של משימות אלו ומציע להפריד אותם כך שכל משימה תילמד בנפרד עד כמה שאפשר. כלומר הלוס יורכב כסכום של שתי פונקציות לוס שכל אחת מהן מתאימה למשימה:

קל לראות ש- היא פונקציית הלוס הקלאסית של סיווג בינארי, כאשר מותאם ללמידה של מודל שפה בלבד. מסמן את הטוקן האמיתי במקום i. שימו לב כי מופיע בביטוי עבור עם stop-gradient. מעצור הגרדיאנט אומר ש- ב- לא משתתף בחישוב הגרדיאנט ולא זורם דרכו. כלומר, מתעדכן רק מ- וכך כל משימה תילמד בנפרד.

הערה: ניתן להתייחס למשימת מזעור פונקציית הלוס כסוג של "תיקון" של סדרת טוקנים הרוסה (corrupted). הרשת מקבלת סדרת טוקנים שיכולה להכיל שגיאות, ומנסה ״לתקן״ אותן ע״י שיוך הסתברות גבוהה לטוקן האמיתי. במאמר, זה נקרא Corrective Language Modeling - CLM. למעשה "CO" הראשון בשם השיטה המוצעת COCO-ML בא מהמילה corrective.

למידה ניגודית על משפטים:

הלמידה הניגודית הפכה להיות גישה מאוד פופולרית בלמידת מפוקחת עצמית (self-supervised learning). המטרה של SSL היא ללמוד ייצוג לטנטי חזק שישמש לאחד מכן עבור משימות למידה שונות. הרעיון העיקרי מאחורי הלמידה הניגודית הוא לבנות זוגות של דוגמאות קרובות (שנקראים זוגות חיוביים) וזוגות של דוגמאות לא דומות, שבדרך כלל נבחרות באקראי מהדאטהסט.

המאמר מציע להשתמש בגישה זו וללמוד ייצוגי משפטים באמצעות הגישה הניגודית. אבל איך נבנה את הזוגות החיוביים והשליליים? הזוגות מורכבים ממשפטים דגומים מהטרנספורמר המשני וחלקים של משפטים אמיתיים (cropped sentences) - פשוט לוקחים תת-סדרה רציפה של הטוקנים מהטקסט האמיתי. הסדרות שנדגמו והוצאו מאותו משפט מהווים זוגות חיוביים ואלו שנלקחו ממשפטים שונים (שנבחרו באקראי) מהווים הזוגות השליליים. פונקציית הלוס הניגודי מוגדרת באופן הבא:

כאשר cos מסמן דמיון מכפלה פנימית (cosine similarity) והטמפרטורה .

פונקציית הלוס של COCO-ML היא למעשה סכום של 3 פונקציות לוס. הראשונה היא הלוס של מודל השפה המשני (auxiliary, שממומש ע״י טרנספורמר) ושני האחרים לוס על הטרנספורמר העיקרי של CLM, והלוס הניגודי .

ביחד:

הישגי המאמר:

לצורך בחינת ביצועי המודל מחברי המאמר פיתחו את מודל COCO-LM תחת שלוש וריאציות שונות של אימון מקדים, הנבדלות ביניהן בכמות הנתונים הזמינים לאימון המקדים וארכיטקטורות שונות של הטרנספורמרים.

המחברים בחנו כל אחד מהמודלים הנ״ל בעזרת מבחני GLUE (קבוצת מבחנים שמשמשים אמת מידה ליכולות של מודלים ב-NLP), והם הושוו למודלים הותיקים RoBERTA ו-ELECTRA, שגם אומנו מחדש על ידי לצורך ההשוואה. בנוסף, המחברים השוו את ביצועי COCO-LM לתוצאות הסטנדרטיות שמקובלות למבחני GLUE.

השוואה למבחני GLUE

בהשוואה של ביצועי מודל COCO-LM, בכל אחת מהוריאציות שלו, ביצועי המודל עלו על ביצועי כל אחד מהמתחרים:

עלות ויעילות האימון (זמן GPU הנדרש לאימון המודל), COCO-LM הצליח להגיע לאחוזי דיוק זהים למודלי RoBERTA ו-ELECTRA תוך שימוש ב-50% ו-60% מזמן ה-GPU, בהתאמה.

כמו כן, בחינה של גרסת ++Large של COCO-LM המודל הציג ביצועים זהים למודל Megatron (בגרסאות של 1.3B ו- 3.9B פרמטרים) במבחן MNLI, תוך שימוש בפחות מ- 10% מכמות הפרמטרים של המודל.

מבחני ablation:

החידושים המרכזיים במאמר היו: הצגה של שתי פונקציות לוס חדשות למודל השפה, פונקציית לוס ללמידה ניגודית של רצפים (Sequence Contrastive Learning, SCL) ולוס ללמידה של מודל שפה מתקן (Corrective Language Modeling, CLM). בחינת ההשפעה של כל אחד מהרכיבים הללו נעשתה על ידי מבחני ablation.

השפעת רכיב הלמידה הניגודית:

לצורך בחינת השפעת רכיב הלמידה הניגודית (SCL) על ביצועי המודל, החוקרים בדקו את השפעת אחוז הקיצוץ (cropping) של הטקסטים וכן את השפעת הלמידה הניגודית על דמיון המשפטים.

החוקרים מצאו שקיצוץ של 10% מהמשפטים שהמודל במהלך הלמידה הניגודית, הביאה לביצועים הטובים ביותר של המודל (שיפור של בין 0.5 ל-0.7 נק׳ בציון הממוצע של מבחני GLUE בהשוואה לקיצוץ של 0% ו 30% בהתאמה).

בנוסף, החוקרים בחנו את ההשפעה של הלמידה הניגודית על ייצוג דמיון בין טקסטים. הדמיון בין טקסטים נמדד בעזרת cosine similarity והוצג בעזרת הורדה של השיכונים המייצגים את הטקסטים למימד נמוך בעזרת t-sne.

כפי שניתן לראות, הלמידה הניגודית הביאה הביאה למבנה יותר טוב של המרחב הלטנטי, שבו טקסטים בעלי דמיון גבוה נוטים להתקבץ לקלאסטרים.

השפעת רכיב למידת מודל שפה מתקן.

ההשפעה של רכיב למידת מודל השפה המתקן (CLM) נעשתה ע״י בחינת דיוק המודל במשימה של תיקון טקסט שהושחת (corrupted), גם באחוז הדיוק כאשר המודל נדרש לתקן טוקנים שהושחתו, וגם בשימור טוקנים קיימים (כאלה שלא הושחתו והמודל נדרש רק להעתיק אותם למיקום המתאים).

רכיב ה-CLM הראה שיפור הן ביכולת לתקן טוקנים שהושחתו, והן ביכולתו לשמר טוקנים, בהשוואה למודל מיסוך סטנדרטי (MLM).

סיכום:

המאמר מציג מודל שפה חדש בשם COCO-LM, שעושה שימוש במודל שפה מתקנת ובלמידה ניגודית. המחברים מאמנים מראש מודלים על רצפי טקסט פגומים, וכך מראים ש-COCO-LM עולה בביצועיו על השיטות הקיימות במדדים של GLUE, תוך כדי ניצול יעיל יותר של משאבי מחשוב ופרמטרי רשת.

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