LLM2Vec: Large Language Models Are Secretly Powerful Text Encoder

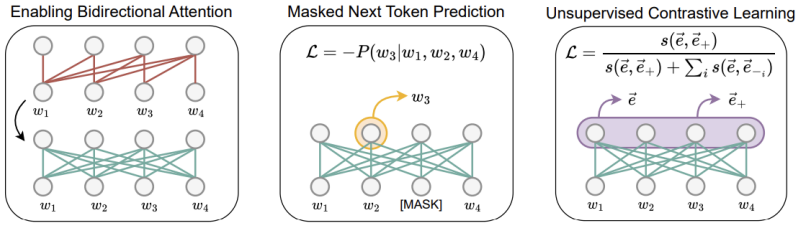


Figure 1: The 3 steps of LLM2Vec. First, we enable bidirectional attention to overcome the restrictions of causal attention **(Bi)**. Second, we adapt the model to use bidirectional attention by masked next token prediction training **(MNTP)**. Third, we apply unsupervised contrastive learning with mean pooling to learn better sequence representations **(SimCSE)**.

המאמר הזה תפס את תשומת ליבו עקב העובדה שהוא דן בנושא שמאוד מעניין אותי לאחרונה (בנוסף לממבה וחידושים למודלי דיפוזיה🙂). והנושא הזה הוא התאמת מודלי שפה מאומנים לביצוע משימות דיסקרימינטיביות, למשל משימות זיהוי נושא או סנטימנט, זיהוי חלקי דיבור וכדומה. הרי רוב מודלי שפה בתקופה האחרונה מאומנים לגנרט טקסט, כלומר לבצע משימה גנרטיבית(מבוססים על דקודר בלבד).

אתם יכולים להגיד למה צריך מודלים למשימות דיסקרימינטיביות אם ניתן די בקלות להפוך רוב המשימות דיסקרימינטיביות לגנרטיביות? למשל משימת זיהוי סנטימנט ניתן להחליף במשימת גנרטיבית של גנרוט הסנטימנט לטקסט נתון (כלומר "הסנטימנט בטקסט זה היה חיובי"). אבל נשאלת השאלה עם החלפה זו היא אופטימלית מבחינת הגודל, הביצועים והמאמץ הנדרש לאימון מודל כזה למשימה נתונה. בלא מעט מקרים (למשל כאשר יש דרישות קשיחות לצריכת זיכרון או לייטנסי מקסימלי של המודל).

האם אפשר לעשות יותר טוב? כאמור רוב המודלים החזקים שיצאו ב-3 השנים האחרונות הם מודלים גנרטיביים בעלי ארכיטקטורת הדקודר (gpt, gemini, claude etc). המודלים שאומנו למשימות דיסקרימינטיביות בעלי ארכיטקטורה הכוללת אנקודר הפכו להיות די נדירים לאחרונה. לאור זה המאמר שנסקורו היום מנסה להתאים (לכייל) מודל שפה גנרטיבי (דקודר) למשימות דיסקרימינטיביות.

עכשיו נשאלת השאלה למה לא לקחת מודל שאומן כדקודר וישר לעשות לו פיינטיון (fine-tune) למשימה דיסקרימינטיבית? כדי להבין למה זה עלול להיות לא אופטימלי צריך להרחיב טיפה על איך בדיוק מאמנים מודלי אנקודר ומודלי דקודר.

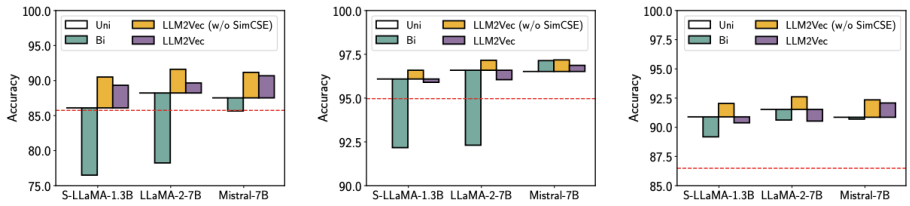
במהלך אימון האנקודר אנו ממסכים טוקנים מסוימים ומאמנים את המודל לחזות אותם. כלומר אנחנו משתמשים **בכל הטוקנים בטקסט** כדי לחזות את הטוקנים הממוסכי. אם הדאטהסט שאנו מאמנים עליו גדול ומגוון מספיק המודל לומד "להבין" (לאפיין סטטיסטית) את השפה. לעומת זאת מודל הדקודר הינו מודל גנרטיבי כלומר המודל יוצר פיסות דאטה חדשות. זה מצריך אופן אימון שונה מהאנקודר. הדקודר מאומן לגנרט דאטה חדש: המודל מאומן לחזות את המילה (טוקן) הבא. כלומר להבדיל מאופן אימון האנקודר אנו **מסתירים מהמודל את הטוקנים שבאים אחרי הטוקן הנחזה, כלומר חוסמים ממנו את העתיד**.

מכאן ניתן לראות עקב אופן אימון שונה קשה וקצת נאיבי לצפות מהמודלים שמאומנים כדקודרים להצטיין במשימות דיסקרימנטיביות אחרי פיינטיון (אני לא טוען שזה בלתי אפשרי וכנראה יש משימות שזה יעבוד להם לא רע, כמובן זה תלוי בכמה דאטה מתיוג יש). נגיד למשימה זיהוי של חלקי דיבור הייצוג של מילה במודל הדקודר המאומן (pretrained) לוקח בחשבון רק את המילים הקודמות שכמובן לא אופטימלי עבור משימה זו.

אחרי הקדמה ארוכה זו בוא נתמקד במאמר המסוקר. כאמור הוא מציע דרך להתאים מודל דקודר מאומן למשימות דיסקרימנטיביות. המאמר מציע 3 שלבים ל"הפיכה" של מודל דקודר למודל האנקודר:

1. ביטול איפוס הטוקנים העתידיים במנגנון ה-attention כלומר המודל חופשי לנצל את כל הטוקנים לבניית ייצוג של כל טוקן . ד"א המאמר טוען הביצועים של המודל לאחר מכן יורדים (בגלל זה יש עוד 2 שלבים בתהליך).
2. במהלך האימון במקום לחזות את הטוקן הממוסך מייצוגו ההקשרי (contextualized) אנו עושים זאת מייצוגו של הטוקן הקודם. לא ברור לי ב 100% מה ההיגיון מאחורי זה.
3. שימוש בלמידה ניגודית (contrastive learning). גישות למידה ניגודית משמשות לאימון של ייצוג דאטה (לא מתויג בד"כ) כאשר מטרת האימון לקרב ייצוגים של פיסות דאטה קרובות ולהרחיק ייצוגים של פיסות דאטה לא דומות/לא קשורות (מבחינת דמיון קוסיין). אז המאמר מציע לאמן את המודל לקרב ייצוגים של אותו המשפט עם drop-outs שונים (בגדול מאוד dropout הוא למעשה איפוס קשרים/משקלים בין נוירונים שונים במודל. לעומת זאת ייצוגים של משפטים מאומנים להיות רחוקים אחד מהם במרחב אמבדינג.

לטענת שילוב שלבים אלו הופך את המודל שלכם לאנדוקר המסוגל להפיק ייצוגים דאטה חזקים המפגינים ביצועים לא רעים בכמה משימות דיסקרימינטיביות.



(a) Chunking (b) NER (c) POS

Figure 2: Evaluation of LLM2Vec-transformed models on word-level tasks. Solid and dashed horizontal lines show the performance of `Uni` and `DeBERTa-v3-large`, respectively.