# Review 92: Revisiting Simple Neural Probabilistic Language Models

**Paper: https://arxiv.org/abs/2104.03474v1**

המאמר עוסק במודלי שפה, שחשיבותם ברורה לכולנו. מודלי השפה הטובים ביותר היום מבוססים על טרנספורמרים. המחברים לוקחים צעד אחורה וחוזרים למודל השפה מבוסס רשתות נוירונים הראשון שהוצע על ידי בנג׳יו ב-2003. השאלה שהם מעוניינים לבחון היא מה יהיו ביצועי המודל כאשר יותאם לתנאים הקיימים כיום - כמות דטה, חומרה מודרנית, שיטות אופטימיזציה מתקדמות ושיטות כלליות כמו layer normalization ו-residual connections.

המודל הבסיסי שהציע בנג׳יו (NPLM) מבצע שרשור של k ייצוגים וקטוריים (embeddings) של המילים שקדמו למילה שנרצה לחזות. הייצוג המשורשר מועבר לרשת feed forward כדי לחזות את המילה הבאה. ההגבלות של מודל פשוט זה הן: 1) גודל החלון שעל בסיסו חוזים את המילה מוגבל (במיוחד אם אמצעי המחשוב מוגבלים) 2) יש לו מספר קטן של פרמטרים מה שמגביל את מידת האקספרסיביות שלו 3) משתמש בסט שונה של פרמטרים התלוי במיקום המילה בחלון.

החוקרים ביצעו מספר שינויים לרשת ובחנו את הביצועים שלה עבור כל סוג שינוי על ידי השוואת ה-perplexity על הדטה סט WIKITEXT-103. השינויים שנבחנו הם הבאים:

הגדלת עומק הרשת וממדי הייצוג הווקטורי של המילים - מ-32 מיליון פרמטרים ל-148 מיליון (כמו הטרנספורמר שהשוו אליו) נצפתה ירידה משמעותית ב-perplexity.

שיפור האופטימיזציה - כדי לשפר את תהליך הלמידה הוסיפו residual connections וlayer- normalization לכל שכבה. כמו כן הוסיפו dropout והשתמשו ב-ADAM optimizer. נצפתה ירידה \*מאוד משמעותית\* ב-perplexity.

הגדלת גודל החלון של המילים הקודמות שעליהן מסתכלים - עד ל-50. נצפתה ירידה ב-perplexity עבור k=3 והגעה לביצועים קבועים בסביבות 40 מילים.

שימוש ב-adaptive softmax וב-tie token embeddings כדי להאיץ את האימון ולהוריד את כמות הזיכרון הדרושה.

הוספת אינפורמציה לגבי הקונטקסט (לא רק k המילים הקודמות) - על ידי הוספה של קונבולוציה נלמדת (חד מימדית) שפועלת על הייצוגים הוקטורים של הקונטקסט.

החוקרים הראו שכל השינוים הללו מביאים לירידה משמעותית מאוד ב- perplexity מ-216 ל- 31.7. טרנספורמר מקבל על אותו הדטה סט תוצאה טובה יותר ב-6 נקודות אבל בכל זאת שמו לב שכאשר הקונטקסט שמשתמשים בו כקלט קטן (כ-3 מילים) ביצועי המודל הבסיסי טובים יותר מביצועי הטרנספורמר. בהשראת התוצאה הזו הם הציעו שני וריאנטים של הטרנספורמר בשניהם שונתה השכבה הראשונה בלבד ושאר השכבות נשארו זהות למודל המקורי.

בווריאציה הראשונה בלוק ה-attention הראשון פשוט מוחלף במודל השפה הפשוט המקבל כקלט חלון בגדול קבוע. בווריאציה השניה הם מגבילים את חלון ה-attention להיות בגודל 5. את הניסויים ביצעו על מספר דטה סטים ונמצא שהוריאנט הראשון נותן את שיפור הביצועים הגדול ביותר לעומת הבייסליין. עבור הוריאנט השני מצאו שככל שגודל החלון קטן יותר שיפור הביצועים גדול יותר.

לבסוף בדקו מה קורה במשימות שבהן פרדיקציה נכונה אפשרית רק כאשר מסתכלים על קונטקסט ארוך. כצפוי ביצועי המודל הבסיסי גרועים מאוד אבל הוריאנטים שהוצעו מראים ביצועים \*טובים יותר\* מאלו של הטרנספורמר המקורי. הם הראו שהביצועים השתפרו עבור מילים שהופיעו יותר מפעמיים בתחילת המשפט, מילים נדירות ישויות.

המאמר הזה נחמד בעיני כי פרספקטיבה היא בסה״כ דבר חשוב. מעניין להבין מה בעצם תורם לביצועים של מודלי השפה הנוכחיים ומפתיע לגלות שגם מודלים ״ישנים״ יכולים להשתדרג לביצועים לא רעים. פחות אהבתי שאין מספיק אינטואיציה לגבי התוצאות. היה נחמד לבדוק למשל שימוש במודל למשימות לינגויסטיות אמיתיות ולראות הבדל בביצועים בין הטרנספורמר ובין המודל שהציעו. גם לא כל כך ברור למה המודל חוזה מילים נדירות באופן מוצלח יותר.

https://arxiv.org/abs/2111.05803