# Review 137: YaRN: Efficient Context Window Extension of Large Language Models, 04.09.2023 https://huggingface.co/papers/2309.00071

**Paper: https://arxiv.org/abs/2309.00071v2**

מכירים את בעיית הרחבת אורך ההקשר (CL=context length) של מודלי שפה? בגדול הסוגיה כאן איך להתאים מודל שפה לעבוד עם טקסטים ארוכים יותר מאלו שהוא אומן עליהם. בהקשר זה קיימות 2 בעיות שונות: איך לגרום למודל שפה לעבוד עם CL ארוך יותר ואיך לכיילו ל-CL ארוך יותר?  
  
אז היום ב-#shorthebrewpapereviews סוקרים מאמר המציע כמה גישות לפתרון של סוגיה זו. המאמר מציע לגשת לפתרון דרך התאמה של קידוד תלוי מיקום (positional encoding = PE) שבא לקודד מיקום של טוקנים בסדרת הפלט. המאמר מציע לשכלל שיטת PE הנקרא RoPE שמקודדת את מיקום התוכן יחד עם ייצוג הטוקן (embedding) כאשר כל מיקום מקודד עם וקטור של רכיבו הוא אקספוננט מרוכב (עם i בתוכו) מוכפל (איבר איבר) עם רכיבי וקטור ייצוג הטוקן עצמו.   
  
כלומר לכל רכיב של וקטור ייצוג התוכן מוספים פאזה (הזזה) שהיא תלויה במיקום הטוקן בסדרה וגם במיקום הרכיב בווקטור הייצוג. קודם כל המחברים שמו לב שבעיית ייצוג המיקום קשורה לתורת NTK (Neural Tangent Kernel)). תורה זו אומרת כי רשתות נוירונים מתקשות לקודד מידע בעלת תדרים גבוהים (כמו קצוות מובהקים בתמונות) עם הייצוג הקלט הוא בעל מימד נמוך יחסית וייצוגו לא מכיל תדרים גבוהים. במודלי שפה אנו מנסים לקודד את מיקום התוכן בסדרה עם וקטור מרוכב(סינוסים וקוסינוסים).   
  
המאמר מדגיש שהגישה הטריויאלית להרחבת חלון ההקשר על ידי הכפלה של כל התדרים של וקטורי הייצוג ביחס בין חלון ההקשר החדש לישן (היחס הזה גדול מ 1) גורמת להפסד של תדרים גבוהים שהרשת צריכה בשביל להבחין בין טוקנים קרובים בעלי ייצוגים קרובים.   
  
בגדול (מאוד) המאמר מציע לא למתוח את כל התדרים במידה שווה (הכפלה ביחס בין אורכי ההקשר) אלא לעשות זאת רק לתדרים נמוכים יחסית (טוקנים ראשונים) ולא לעשות זאת (הכפלה) לטוקנים בסוף הסדרה כאשר טוקנים באמצע נמתחים גם כן אבל עם מקדם יותר קטן מהיחס הזה. כמובן שזה ממש בגדול, מציע בחום לרקוא את המאמר הזה כי אני מרגיש שהוא עוד יעשה כותרות.