# Review 134: LM-INFINITE: SIMPLE ON-THE-FLY LENGTH GENERALIZATION FOR LARGE LANGUAGE MODELS, 31.08.23 https://arxiv.org/abs/2308.16137.pdf

**Paper: https://arxiv.org/abs/2305.13281v1**

אורך ההקשר (context length) של מודלי שפה או במילים אחרות הגודל המקסימלי של הטקסט המודל שפה יכול ״לזכור הפך להיות נושא מאוד פופולרי בקהילת ה-NLP עקב חשיבותו הרבה ליישומים רבים. יצאו עשרות (אם לא מאות מאמרים) המנסים להגדיל אותו והיום ב-#shorthebrewpapereviews אנו נסקור מאמר המציע גישה חדשה להארכת אורך הקשר.   
  
קודם כל המחברים מזהים (וסוג של מוכיחים) למה מודלי שפה שאומנו עם קלט קצר יחסית ועם קידוד תלי מיקום יחסיים (RoPE) מתקשים עם קלט יותר ארוך באינפרנס. אז לפי המאמר יש 3 סיבות עיקריות.

או ש- rope מתעלם מהטוקנים הרחוקים (מקדמי attention לפני softmax שווים ל-0 ) או שהם מקבלים ערכים גבוהים מאוד

ככל שמאריכים את אורך ההקשר האנטרופיה של מקדמי ה-attention שואפת ל-log(N) כאשר N זה מספר הטוקנים כלומר המודל מתחשב בכל הטוקנים באותה מידה

המודל מקודד באופן לא מפורש את המיקום האבסולוטי של הטוקנים בסדרה (המאמר מסביר את זה בכך ש״הסיגנל מהטוקנים ההתחלתיים חזק יותר מהאלה שבסוף)  
  
כדי להתמודד עם 3 סוגיות האלה המאמר מציע גישה די פשוטה לקידוד תלוי מיקום:

עבור טוקן נתון הם מקודדים (עם RoPE או משהו דומה) מספר טוקנים מסוים (נגיד שווה לאורך ההקשר ״הסטנדרטי״ של מודל שפה) המופיעים בתחילת הסדרה (global branch)

כל טוקן גם מקודד (attend) את מיקום טוקנים הנמצאים במרחק מסוים ממנו (local branch)

לפי המאמר כך הטוקנים שבהתחלה בעיקר מקודדים בעיקר את המיקום האבסולוטי של הטוקן, אלה שבסוף את המיקום היחסי שלו ואלו שבאמצע מכילים פחות מידע מיקומי ( לא הבנתי את למה בעצם).