# Review 152: CoCA: Fusing Position Embedding with Collinear Constrained Attention in Transformers for Long Context Window Extending, 23.09.2023

**Paper: https://arxiv.org/abs/2309.08646v3**

https://arxiv.org/abs/2309.08646.pdf

זה לא סוד שאחד המרכיבים החשובים בטרנספורמרים הינו קידוד תלוי המיקום (positional encoding) או PE. תפקיד PE הוא לקודד מיקום הטוקנים בסדרה והשיטת קידוד שהפכה להיות מאוד פופולרית לאחרונה נקראת (RoPE (rotary PE.   
  
אז היום ב-#shorthebrewpapereviews אנו סוקרים מאמר שמצא מקום בו ניתן לשפר (לפי המאמר) את RoPE ומציע דרך לפתור אותה. אז קודם כל מה זה RoPE? זו שיטה שלמעשה לוקחת וקטורי שאילתה ומפתח (query and key) ומכפילה אותם (איבר איבר) בווקטור מרוכב בעל נורמה יחידה שהתדר שלו פרופורציונלי למיקום של טוקן בסדרה (כל איבר בווקטור זה מוכפל גם במימד שלו במרחב הייצוג).   
  
כלומר ככל שהטוקן נמצא יותר רחוק מתחילת הסדרה התדר שלו (מקדם מעריכי במספר המרוכב הזה) הינו גבוה יותר. צריך לציין שוקטורי המפתח והשאילתה מיוצגים כוקטורים מרוכבים גם כן. כאשר מחשבים את ה-attention בין וקטורים אלו יוצא כי יש פונקצית ה-attention תלויה באופן מפורש במרחק בין וקטורים אלו (נמצא בתוך אקספוננטה מרוכבת).   
  
ניתן להוכיח ככל שעבור מרחק גדול בין הטוקנים ה-attention ביניהם שואף לאפס. עכשיו המחברים שמו לב שעבור מימדים מסוימים במרחב הייצוג ציוני ה-attention בין מקדמי שאילתה ומפתח (עבור כל מימד מדובר בשני זוגות של מספרים מרוכבים) עלולים לקטון כאשר מרחק בין מיקומי הטוקנים קטן (בגלל המבנה של RoPE). כמובן שזה לא רצוי ולמרות שזה קורה רק למימדים מסוימת המחברים מוכיחים שזה משפיע לרעה על יעילות הקידוד המיקומי.   
  
הסיבה לכך (הנובעת מאריתמטיקה די פשוטה) היא הזווית שהיא לא אפס בין וקטורי השאילתה והמפתח. אז המחברים מציעים שיטה ההופכת וקטורי שאילתה להיות קולינאריים כלומר הזווית ביניהם הופכת להיות 0 והבעיה נעלמת. יש שיפור מסוים בביצועים אך המחברים עצמם אומרים שטרם סיימו לבדוק את כל ההיבטים של הגישה המוצעת.