# Review 187: Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention

**Paper: https://arxiv.org/abs/2006.16236v3**

https://arxiv.org/pdf/2006.16236

אחרי הסקירה הקודמת הכבדה מאוד מחכה לנו היום סקירה קלילה (הסקירה הבאה הולכת להיות די כבדה). כמו שכבר אמרנו אחד החסרונות הבולטים של הטרנספורמר היא הסיבוכיות הריבועית שלו במונחי אורך הקלט (= מספר איברים בסדרת הקלט). הסיבוכיות הזו בא על ידי ביטוי גם במהלך האימון וגם במהלך ההיסק (inference). סיבוכיות ריבועית זאת כואבת במיוחד בזמן ההיסק כאשר אין לנו יכולת לחזות מספר טוקנים בו זמנית כי לחיזוי טוקן n אנו צריכים לדעת את ה-(n-1) הטוקנים הראשונים. האם ניתן להפוך את הטרנספורמר לסוג של RNN במהלך ההיסק כאשר כל הזיכרון על הטוקנים הקודמים נדחס לכמה וקטורים בודדים (וקטור זכרון ווקטור של המצב)?

הטרנספורמר המקורי אינו מאפשר אופן חישוב כזה כי הוא מכיל פעולה לא לינארית (softmax) בתוך מנגנון תשומת הלב שלו. ניתן לראות די בקלות שלא ניתן לעקוף את מגבלת הסיבוכיות הריבועית שלו ללא שינוי של אופן חישוב של תשומת הלב. המאמר המסוקר מציע להחליף את חישוב הסופטמקס במנגנון זה בחישוב לינארי (מכפלת מטריצות) המחושבות על ידי הפעלת פונקציה לא לינארית phi על וקטורי השאילתות Q ושל וקטורי המפתחות K. מי שעוד זוכר מה זה KT)Kernel Trick) מבין מה שנעשה כאן הוא KT בכיוון ההפוך.

כמובן שאנו מאבדים כאן מהעוצמה של המנגנון תשומת הלב הרגיל אבל זה יעזור לנו לפתור את סוגיית הסיבוכיות הריבועית בזמן ההיסק. למעשה המחברים מוכיחים (ראו את התמונה למעלה) כי ניתן לממש את המנגנון הזה לסדרתי בעל סיבוכיות לינארית במונחי אורך הקלט. כמובן בזמן האימון ניתן לחשב חיזוי של כמה טוקנים בו זמנית (לפי היכולת החישובית שעומדת לרשותנו) וליהנות מהיתרון של מנגנון תשומת הלב הרגיל.

כלומר יש לנו טרנספורמר (מוחלש כמובן) באימון ו- RNN בהיסק. בהמשך נראה ניתן לשפר את הגישה הזו עם SSMs) state-space models).