# Review 109: [Short] Scaling TransNormer to 175 Billion Parameters, 28.07.23 https://huggingface.co/papers/2307.14995

**Paper: https://arxiv.org/abs/2410.21016v2**

אתם יודעים שחישוב ה-attention בתוך הטרנספורמר הוא ריבועי במונחי אורך של הטקסט? יש כמה טריקים כמו FlashAttention שמנצלים את האופיינים של gpu ומצליחות להקטין משמעותית את מס' הפעולות אבל החישוב עדיין נותר די כבד. אבל לא עוד!!

היום ב-#shorthebrewpapereviews סוקרים מאמר המציע מנגנון ה-attention בסיבוכיות לינארית שהוא עוקף את המנגנון המקורי מכל הבחינות! כלומר גם ביצועים טובים יותר במשימות מגוונות וגם זמן ההסקה הממוצע שלו (inference) נמוכים יותר. איך הם עשו זאת?

בגדול (ויש הרבה פרטים קטנים אך חשובים בתוך המאמר עצמו) המאמר מחליף את ה-softmax שמופיע במנגנון תשומת הלב במכפלה של המטריצות שמאפשר להחליף את סדר הכפלת המטריצות שמביא לנו את הסיבוכיות הלינארית הנחשקת מבחינת אורך ההקשר. אבל לא מעט מאמרים ניסו את הגישה הזו אך לא הצליחו אז "the devil is in the details".

אחד הפרטים המעניינים הוא שהמחברים משתמשים בקידוד תלי מיקום יחסי (relative positional encoding) ולא בקידוד הסטנדרטי של הטרנספורמרים. איך מבצעים את הקידוד הזה? במקום לחבר אותם לייצוגי הטוקנים מחברים את הקידוד היחסי הזה למכפלת מטריצות שאילתה והמפתח בתוך מנגנון ה-attention.

שימו לב שעבור שני טוקנים במיקומים s ו- t הקידוד הזה תלוי בטוקנים עצמם ובהפרש המיקומים t-s. במרבית המקרים קידוד זה מכיל חלקים קבועים (כמו דעיכה מעריכית במונחי s-t וגם חלקים נלמדים). אז כאמור בנוסף לכך המאמר מציע מכלול של טריקים חישוביים כדי להגיע לתוצאה המיוחלת: ניצחון על הטרנספורמים הקלאסיים!