⚡️🚀המאמר היומי של מייק -05.11.24: ⚡️🚀

RETHINKING SOFTMAX: SELF-ATTENTION WITH POLYNOMIAL ACTIVATIONS

מאמר די לא רגיל והוא מדבר על חלופה פוטנציאלית של מנגנון ה-attention שאנו כה אוהבים בטרנספורמים. אתם בטח זוכרים שמשקלי attention בשנאים מחושבים עם softmax שהוא מנרמל וקטורי משקלים לנורמה 1 ובנוסף כל רכיביו הינם בין 0 ל- 1 כלומר הוא מהווה התפלגות הסתברותית. המחברים טוענים שתכונות אלו של המשקלים לא קריטיות לפונקציונאליות של השנאים ומציעים להחליף אותם בקרנל אחר שהוא פולינומיאלי כפי שאתם בטח ניחשתם מהשם של המאמר.

אבל למה זה עובד בכלל? המחברים טוענים (באופן די מפתיע, אני חייב להגיד) שהביצועים הנפלאים של הטרנספורמרים נובעים בחלקם מיכולתה של פונקציית סופטמקס לכפות רגולריזציה מסוימת על נורמת פרובניוס של מטריצה המשקלים וגם של היעקוביאן שלה (ביחס לקלט של הסופטמקס) במהלך האימון הוא מסדר (sqrt(n כאשר n הינו מימד לקלט.

נורמת פרובניוס או NF מוגדרת בתור שורש של סכום הריבועים של כל הערכים במטריצה והיא גם שווה לשורש של סכום הריבועים הערכים הסינגולריים (הכללה של ערכים עצמיים למטריצות לא ריבועיות). ד״א סופטמקס מחושב במנגנון ה-attention של מערך של וקטורים אז היעקוביאן תיאורטית הוא טנזור תלת מימדי (המאמר מפרט איך מחשבים את NF במקרה הזה).

אז בגדול המאמר מוכיח שני משפטים. בראשון מהם טוענים ש NF של מנגנון attention פולינומיאלי (כולל הלינארי) מתנהג לפי (O(n אם המטריצות שם, K ו-Q וגם ייצוגי הטוקנים מפולגים גאוסית כמובן). אז אם מנרמלים את ה-attention הפולינומיאלי עם (n^(-0.5 מקבלים את (sqrt(n שהיה לנו עבור מנגנון ה-attention הרגיל. בנוסף NF של היעקוביאן לפי Q, המנורמל לפי (n^(-0.5 (לא זה שמתנהג לפי (sqrt(n ב-attention הרגיל) גם מתנהג לפני (sqrt(n.

המחברים טוענים שזה מספיק כדי לטעון שניתן להחליף סופטמקס בפולינומים שיותר קלים מבחינה חישובית, מקבלים תוצאות מעודדות אבל אני עדיין לא השתכנעתי…

https://arxiv.org/abs/2410.18613