⚡🚀המאמר היומי של מייק 08.06.24:⚡🚀

Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality

למאמר הזה יש עוד שם והוא 🤔mamba-2. המאמר הזה מתמקד בשכלול הארכיטקטורה של ממבה המקורית שעשתה הרבה כותרות בחצי השנה האחרונה ואני הצטרפתי לחגיגה וסקרתי בערך 20 מאמרים בנושא המרתק הזה.

המאמר הזה של Albert Gu התותח ממשיך להעשיר את עולם הממבה והפעם הוא הגיע לכמה תובנות די מעניינות. הוא לראשונה מגדיר SSM בעל תכונה N-semi-separable שלמעשה מגדיר את צורתו של קרנל קונבולוציה המופעל על סדרת הקלט במוד הקונבולוציוני של SSM (כאשר משתמשים ב-SSM לאימון ממוקבל). אלחש לכם בסוד שבסופו של דבר זה מתנקז לצורתו של מטריצה A.

שנית מאמר חוקר מנגנוני ה-attention בפרט השונים למשל הקלאסי הלינארי ,כלומר ללא סופטמקס, ועם סדר שונה בביצוע פעולות בין מטריצות Q, K, ו-V. המאמר מפרק את החישוב ל- 3 שלבים "אטומיים" (שכל אחת מהם הוא מכפלות מטריצות, אך לפעמים מאוד גדולות) השלב השני והחשוב ביותר הוא מיסוך (masking) שניתן לתאר אותו גם עלי ידי מכפלות מטריצות (Kernel trick). המיסוך הקוזלי (causal) הוא חלק ממנגנון ה masking. הבחנה זו אפשרה למחברים להוכיח סוג של שקילות בין מנגנוני attention מסוימים ל-SSMs .

בנוסף הם מפתח שיטה לחישוב יעיל של קונבולוציה ארוכה (שזה הלב של SSM) בחומרה עם מטריצות הנקראת 1-semi-separable (עבור מטריצה A מצורה מסוימת).

מה יוצא לנו מכל הסיפור הזה? האצת אימון של ממבה (שזה למעשה ממבה 2) וגם פריימוורק תיאורטי למידול ארכיטקטורה העוצמתית הזו משותפת גם למנגנוני ה-attention השונים.

קריאה מהנה!

https://arxiv.org/abs/2405.21060