⚡️🚀המאמר היומי של מייק 07.09.24: ⚡️🚀

ReMamba: Equip Mamba with Effective Long-Sequence Modeling

 סוקר את המאמר הזה משתי סיבות. קודם כל הוא קשור לממבה. הסיבה השני היא זה שהתבקשתי לסקור אותו. ואוקיי, המאמר לצערי לא חידש לי הרבה ולדעתי לא נזכור אותו בעוד כמה חודשים.

אתם זוכרים את State Space Models או SSM בהקשר של למודלים עמוקים? SSM הוא ארכיטקטורה יחסית חדשה עבור רשתות לעיבוד דאטה סדרתי (שפה טבעית וגם תמונות). השוס הגדול ב-SSM היא שהם מאוד מהירים גם באימון וגם באינפרנס עקב כך שניתן לייצג אותם בתור רשת קונבולוציה וגם במודל רשת recurrent. הגמישות הזו כמובן גובה מאתנו מחיר בדמות חוסר expressiveness (יכולת למדל חוקיות מורכבות) של ארכיטקטורה הזו עקב העובדה המעברים בין המצבים החבויים הם לינאריים וקבועים לכל איברי הסדרה (מכאן בא הדואליות בייצוג).

ארכיטקטורת ממבה מחזירה לנו קצת מה-expressiveness בכך שהופכת את המעברים בין המצבים החבויים לתלוי במצב החבוי אך משאיר אותם לינאריים. זה עוזר אבל עדיין ממבה מתקשה במשימות reasoning מורכבות עקב מחסור ב-expressiveness. ייתכן שאחת הסיבות לאי הצלחה זו היא חוסר יכולת של ארכיטקטורת ממבה לדחוס את המידע הרלוונטי למשימה (לגיטימי אבל כמובן יש עוד סיבות לכך).

המחברים מציעים לדחוס את ייצוגיהם של תת סדרות של טוקנים. נניח שיש לנו L טוקנים בהקשר ואנו רוצים ״לדחוס״ את טוקנים שייצוגיהם דומה לזה של הטוקן L. כלומר מחשבים את הדמיון בין תת-סדרה רציפה נתונה של טוקנים (הייפרפרמטר) ודוחסים את הייצוגים של הטוקנים בתת-סדרה זו לפחות טוקנים (הייפרפרמטר גם כן). כלומר במקום הייצוגים של M טוקנים בתת סדרה נקבל ייצוגים של K טוקנים אחרי הדחיסה.הדמיון מחושב דרך דמיון קוסיין (עם כל מיני שכבות לינאריות).

הם מראים שזה עובד - לי זה מריח קצת אוברפיט וגם קושי באופטימיזציה של ההייפרפרמטרים….

https://arxiv.org/abs/2408.15496