המאמר היומי של מייק - 12.02.25  
STUFFED MAMBA: State Collapse and State Capacity of RNN-Based Long-Context Modeling

המאמר מספק חקירה מעמיקה של מצבי כשל במודלים מבוססי RNN במידול שפה עם הקשר ארוך ומציע פתרונות לשיפור יכולות ההכללה שלהם לאורכים גדולים. המחברים מזהים ומנתחים תופעה בעייתית מאוד שקיבלה שם קריסת מצב (State Collapse - SC) - כשל של מודל בעקיבה אחרי דינמיקת של הדאט המונע מרשתות RNN להכליל מעבר לאורכי האימון שלהן. הם מציגים סט של טכניקות מיטיגציה ללא אימון ואסטרטגיות אימון המשכי המאפשרות למודל Mamba2 לעבוד עם מעל מיליון טוקנים מבלי לסבול מקריסת מצב.

הגדרת הבעיה:

מודלי RNN לעומת טרנספורמרים במידול הקשר ארוך

- טרנספורמרים משיגים ביצועים עדיפים במשימות המצריכות הקשר ארוך אך סובלים מסיבוכיות חישובית ריבועית ביחס לאורך הסדרה בשל מנגנון attention.

- מודלי RNN מציגות סיבוכיות לינארית ביחס לאורך הסדרה, מה שהופך אותן ליעילות חישובית בטיפול בסדרות ארוכות.

- מודלים בעלות סיבוכיות לינארית כמו Mamba, RWKV מאומנים בד״כ על סדרות קצרים יחסית (~10K טוקנים) ונכשלים בהכללה מעבר לאורכי האימון(זו הטענה במאמר)

ניתוח כשלים ברשתות RNN (וגם Mamba, RWKV) עם הקשר ארוך

כישלון בהכללה עבור סדרות ארוכות יותר: רשתות אלו מציגות הידרדרות חדה בביצועים כאשר נחשפות לאורכי סדרות מעבר לדאטה שאומנו עליו. כישלון זה אינו נובע פשוט מגרדיאנטים דועכים אלא מיוחס לקריסת מצב (SC).

קיבולת זיכרון קבועה: מכיוון RNNs שומרות על מצב זיכרון בגודל קבוע, יכולתן לשמור מידע היא מוגבלת מטבעה. קיימת מגבלה עליונה על קיבולת הזיכרון ההקשרי - טוקנים מעבר למגבלה זו נשכחים בהכרח.

2. ניתוח פורמלי של קריסת מצב (SC)

הגדרה וממצאים: קריסת מצב (SC) מתרחשת כאשר התפלגות המצב החבוי קורסת(מתנוונת), מה שמוביל לכישלון המודל בעיבוד רצפים ארוכים יותר מקבוצת האימון. המחברים מבצעים ניסויים מבוקרים על Mamba2 ומבחינים שערוצי מצב חבוי מסוימים מציגים התפוצצות של שונות, הגורמת ל:

- ערוצים(channels) חריגים דומיננטיים המדכאים ערכי מצב אחרים.

- חוסר יכולת לשכוח טוקנים מוקדמים, המוביל זיכרון.

- SC מתבטא בעלייה חדה בפרפלקסיות(אי וודאות) מעבר לאורך האימון.

ייחוס תיאורטי: פרמטריזציית יתר בדינמיקת המצב

המחברים מנסחים את משוואת עדכון המצב:

כאשר h\_t הוא וקטור המצב החבוי, המקדמים מייצג את קצב דעיכת הזיכרון, מייצג מידע חדש שהוכנס בצעד זמן. כאשר מאמנים על סדרות באורך פרמטרי המודל הנלמדים מעדיפים ״לשמור את כל המידע בתוך,״ ועקב כך נכשלים בעת עיבוד סדרות ארוכות יותר. זה מוביל לצבירת יתר של מידע, שמובילה לרוויה ובסופו של דבר לקריסת מצב.

3. אסטרטגיות התמודדות נגד SC:

טכניקות הת ללא אימון של SC:

שכחה מבוקרת: הגדלת דעיכת ייצוג מצב(חבוי) על ידי שינוי גורם הדעיכה והפחתת ״עוצמת הכנסה״ של מידע חדש (ייצוג של טוקן) . צעדים אלו גורמים למודל לשכוח טוקנים ישנים באופן אפקטיבי, מונע מייצוג הזכרון להגיע לרוויה (ערכים גבוהים מדי).

נרמול מצב: החלת אילוץ מבוסס נורמה על ייצוג המצב החבוי (מחלקים את וקטור הייצוג בנורמה שלו אם היא גדולה מדי):

זה מונע התפוצצויות של ייצוג מצב חבוי אך מכניס אי-לינאריות, המשפיעה על יעילות האימון (לא ניתן למקבל את החישובים).

עדכון וקטור ייצוג המצב עם sliding window: ניסוח מחדש של כלל עדכון ייצוג המצב לסימולציה של מנגנון sliding window\*\*:

זה מסיר טוקנים ישנים באופן אפקטיבי מבלי לחשב מחדש מאפס. ישים לארכיטקטורות אחרות כמו RWKV ו-RetNet.

המשך אימון על רצפים ארוכים יותר: המחברים מרחיבים את אורכי דאטה האימון מעבר ל״קיבולת ייצוג המצב״ כדי לאלץ את המודל ללמוד כיצד לשכוח בהדרגה. הם מאמתים אמפירית שעבור כל גודל ייצוג מצב S, קיים סף אורך אימון שבו SC לא מתרחש.

4. סיכום:

- המחקר השיטתי הראשון של קריסת ייצוג מצב (SC) ברשתות ״דמויות״ RNN עם אורך הקשר ארוך.SC מתבטא בכך שוקטור ייצוג המצב מגיע לרוויה (ערכים גבוהים) וזה גורם להידרדרות רצינית בביצועי המודל. המאמר מציע 3 שיטות מיטיגציה ללא אימון לביטול SC עד מיליון טוקנים. המחברים הציעו ביסוס אמפירי לקשר בין \*\*גודל ייצוג המצב לקיבולת המודל. לבסוף הם אימנו מודל Mamba2 בעל 370M פרמטרים עם אחזור מושלם של 256K טוקנים - הרבה מעבר ליכולות של מודל סטנדרטי מסוג זה.

https://arxiv.org/abs/2410.07145