# Review 188: Efficiently Modeling Long Sequences with Structured State Spaces

**Paper: https://arxiv.org/abs/2111.00396v3**

https://arxiv.org/abs/2111.00396

לאט לאט הגענו למאמר הרביעי בסדרת סקירות בדרך לממבה. הפעם נסקור מאמר מ-2022 שיצא שנתיים אחרי 3 המאמרים הראשוניים שסקרנו בנושא המעניין הזה. כמובן במהלך תקופה זו יצאו כמה מאמרים מעניינים שפיתחו ארכיטקטורות מבוססות מערכות דינמיות לינאריות (ובשם כללי יותר Space-State Models- SSMs).

המאמר שנסקור לקח את הגישה הזו לגבהים חדשים והגיע לתוצאות די מרשימות עם דאטה בעל אורך הקשר ארוך (למשל עבור אות אודיו המכיל אלפי או אפילו עשרות אלפי דגימות בשנייה. אם יש לנו מטלה שדורשת התחשבות בכמה עשרות שניות של אודיו אז אנו צריכים אורך הקשר של מאות רבות של דגימות וזה די כבד עבור הטרנספורמר עם הסיבוכיות הריבועית שלו - במונחי אורך הקשר).

אוקיי, אז בואו ניזכר מהו היתרון הבולט של ארכיטקטורות מבוססות SSMs. מצד אחד בעת ההיסק (inference) של טוקן הם מונעים מאיתנו צורך להתחשב באופן מפורש בכל הדגימות הקודמות על ידי דחיסה של המידע בטוקנים הקודמים(=זיכרון) בווקטור זיכרון אחד, המתעדכן עם המערכת הדינמית הלינארית. מצד שני במהלך האימון (כשכל הטוקנים ידועים) הוא מאפשר חישוב בו זמני של כל הטוקנים הממוסכים.

דואליות עוצמתית זו התאפשרה על ידי ייצוגה של זיכרון בתור מערכת לינארית שניתן לבטא את הזיכרון המצטבר לכל טוקן כפעולה לינארית. כלומר ניתן לתאר את הפלט של עבור טוקן k על ידי הנוסחה באחת התמונות (הקטנה יותר).

מטריצות בנוסחה הן הגרסאות המודסקרטות של המטריצות המופיעות בנוסחה של המערכת הדינמית המתארת את התקדמות הזכרון בזמן (טוקנים). ניתן לראות כי מה שיש לנו כאן זו רשת קונבולוציה(שעלולה להיות מאוד ארוכה) שמאפשרת חישוב הייצוג של כל טוקן i.

קיבלנו את הארכיטקטורה הדואלית המתאימה גם לאימון וגם להיסק. אבל יש בעיה קטנה. עבור אורך הקשר גדול מספיק נדרשת כמות גדולה מאוד של זכרון. קודם כל אנו צריכים מטריצה A בגודל NxN (נגיד עבור N=64) עבור כל מימד של ייצוג הקלט (כי זה מה שהמערכת הדינמית שלנו ״צריכה לזכור״). אז חישוב קונבולוציה זו בצורה הישירה עבור מטריצה A כללית של המאמר של HiPPO (עבור מקרה של פולינומי Legendre שנקרא LegT תחת מכסה המנוע של המערכת הדינמית) הוא מאוד כבד ודורש הרבה זיכרון.

אז מה ניתן לעשות? קודם כל אם מטריצה A היא אלכסונית החישוב ודרישות הזכרון היו הרבה יותר צנועות. המחברים גם שמו לב כי conjugation של מטריצה A במערכת הדינמית (הכפלתה מימין ומשמאל במטריצה אוניטרית V) מוביל למערכת דינמית שקולה עם התוצאה Vx. הבעיה שמטריצה A מ-HiPPO לא ניתן לתאר בצורה V\*LV כאשר L היא מטריצה אלכסונית, ו V היא מטריצה אוניטרית (נובע מכך ש A אינה קומוטטיבית עם A\* כלומר לא נורמלית - זה השם אין מה לעשות).

אז הכל אבוד? מתברר שלא. מתברר ש A מ HiPPO ניתן לתאר בתור סכום של מטריצה נורמלית ומטריצה בעלת רנק נמוך (עבור LegT הרנק אפילו שווה ל-1 כלומר תוספת זו כי מכפלה חיצונית של שני וקטורים בעלי מימד Nx1). ואז המאמר מציע אלגוריתם די לא טריוויאלי עבוד חישוב של קרנל קונבולוציה ארוך המבוסס על 3 עקרונות מתמטיים:

- במקום לחשב A^l עבור כל l ניתן לחשב z-transform (מקוטע עד L) של A ואז לחשב בצורה די פשוטה את A^l על ידי הצבה של שורש שונים של 1 (המרוכבים) ב z-transform הזה.

- כאשר A הוא הפרש של מטריצה אלכסונית L ומטריצה בעלת רנק נמוך מאוד ניתן לחשב את z-transform הזה בצורה יעילה דרך זהות Woodbury שמסתכם בהיפוך של מטריצה אלכסונית.

- ניתן לבצע את כל החישובים העולים כאשר מפעילים בזהות Woodbury בצורה יעילה מאוד עם Cauchy Kernel; שזה בגדול מטריצה שנבנית בצורה מסוימת משני וקטורים

לבסוף, מבצעים את החישובים האלו עבור כל מימד של ייצוגי הטוקנים בנפרד ואז מערבבים עם שכבה לינארית (או כמה). מטריצות אלכסוניות L (למעשה וקטור), וקטורים B, C וגם P ו-Q שמכפלתם היא מטריצה בעלת נמוך מאומנות בנפרד עבור כל מימד של ייצוג הטוקנים.

זהו, יצא ארוך - הסקירה הסקירה תהיה קצרה יותר.