# Review 194: Mamba: Linear-Time Sequence Modeling with Selective State Spaces

**Paper: https://arxiv.org/abs/2507.06204v1**

https://arxiv.org/abs/2312.00752

זה קורה עכשיו, אחרי 9 סקירות שחלקם היו די לא פשוטות הגענו למטרתנו הקדושה שזה Mamba. מכיוון שאני מפרסם סקירות בשלשות (באתר MDLI) אני אוציא עוד 2 סקירות נוספות של שכלולי ממבה (אחד מהם Mamba MoE והשני עוד הוחלט).

האמת שאחרי שאנו הבנו מה- (SSM (space-state models ואיך ניתן לבנות ארכיטקטורה מבוססת עליהם לעיבוד דאטה סדרתי, השכלול המוצע על ידי mamba הוא די אינטואיטיבי ומתבקש. כמו שאתם זוכרים SSM ממומשת בתור מערכת דינמית(DLS) לינארית כאשר הקלט למערכת זו היא ייצוג וקטורי (embeddings) של איברי הסדרה (= טוקנים).

בשלב הראשון המערכת הדינמית מחשבת וקטור s הוא הוא ייצוג דחוס של זיכרון כלומר וקטור ״הזוכר״ את המידע הרלוונטי עבור כל הטוקנים הקודמים לטוקן הנוכחי. בשלב השני מחשבים את הפלט עבור טוקן זה המוזן לשכבה הבאה (שיכולה להיות גם שכבת שמייצרת פלט סופי). כל חישובים אלו מתבצעים באמצעות מיפויים לינאריים כלומר מכפלות במטריצות. חשוב להבין שכל המעברים בין ייצוגי הזכרון בין הטוקנים הם לינאריים ונשלטים על ידי אותה מטריצה A ווקטורי B, C וסקלר delta. ד״א פרמטר delta מגדיר (באופן גס) את קצב דעיכה של הזכרון (כלומר ככל ש delta גבוה יותר אנו נוטים ״לזכור״ פחות מהטוקנים הקודמים).

מה היתרונות של הארכיטקטורה הזו? היא בעלת תכונה הדואליות המיוחלת המשלבת 2 התכונות הבאות:

ניתן לחזות באופן מקבילי (בו זמנית) כמה טוקנים במהלך אימון (כמו בטרנספורמרים)

חיזוי מהיר של טוקן במהלך היסק (ללא התחשבות בכל הטוקנים בחלון ההקשר כמו בטרנספורמרים שמביא לנו את הסיבוכיות הריבועית).

כלומר הארכיטקטורות מסוג זה הם יעילות בזמן האימון ומהירות בזמן ההיסק. אבל כמו שאתם יכולים לנחש יש לנו מחיר לשלם על כל התכונות הנחמדות האלו. ומחיר הוא כמובן יכולת של המודל למדל תלויות מורכבות של הדאטה. עקב כך מאמרים כמו Hyena, H3, S4 ניסו ניסו לבנות את הפרמטרים של DLS (המגדירה מעברים בין ייצוגי הזכרון ויצירת הפלט) בצורה חכמה (ודי מורכבת).

אבל מתברר שזה לא מספיק. מעברים לינאריים עם פרמטרי DLS קבועים לא מסוגלת למדל דאטה מורכב (כמו שפה טבעית). אחד המשימות שמודל כזה נכשל עליה הוא העתקת טוקנים הבאים אחרי טוקן ספציפי (וזה די הגיוני לאור הפרמטרים הקבועים של DLS). כמו שאתם יכולים כבר לנחש אולי מחברי ממבה מציעים לעשות חלק מהפרמטרים (B, C ו delta) תלוים בייצוג הטוקן הנוכחי. התלות הזו היא לינארית עם מטריצות נלמדות. וזה עוזר לנו להתחשב בפיסת הקלט הנוכחית בצורה יותר טובה. כאמור B, C מגדירות את האופן בו ייצוג הזכרון והפלט עבור הטוקן הנוכחי בהתאמה ואז יש לנו סיכוי יותר טוב להצליח במשימות מהסוג שתיארתי לפני. בנוסף תלות של delta בייצוג הטוקן הנוכחי מקנה לנו אפשרות לשחק עם קצב דעיכה בצורה יותר גרנולרית שמקנה לנו יכולת ״לשכוח״ ו״לזכור״ איפה שצריך.

אבל האם איבדנו את הדואליות שלנו בדרך. מתברר שלא, הרי המעבר בין ייצוגי הזיכרון של הטוקנים עדיין לא תלוי במיקום של הטוקן אלא בייצוג. כלומר אנו עדיין יכולים לחזות מספר טוקנים בו זמנית כי אנו יכולים לחשב את כל הפרמטרים מראש (לא צריך לחשב את המצב הזכרון הקודם באופן מפורש). וכמובן אין צורך להתחשב בכל הטוקנים בתוך חלון ההקשר במהלך ההיסק כי הזיכרון עדיין מיוצג על יד וקטור אחר. אז יש דואליות!

מה שכן קורה הוא זה החישובים הופכים לקצת יותר מורכבים (שימו לב שהחלק הכי בעייתי בחישוב שהוא העלאה של מטריצה בחזקה לא השתנה כי A נותרה קבועה). בסוף המאמר מציע כמה שכלולים לאופן חישוב המייעלים ומזרזים אותו (המשחקים בין זכרון מהיר ואיטי של GPU).

זה וזה עכשיו אתם יודעים מה זה ממבה. נתראה ב Mamba MoE עוד כמה ימים.