# Review 196: VMamba: Visual State Space Model

**Paper: https://arxiv.org/abs/2401.10166v4**

https://arxiv.org/abs/2401.10166

מתחילים את הסקירה האחרונה בסדרת מאמרי ממבה. באופן די טבעי המאמר הזה מדבר על שילוב של ארכיטקטורה זה למודלי הראייה הממוחשבת (או ויז'ן בקצרה). הסקירה הולכת להיות די קצרה וקלילה.

הפעם לא אספק לכם סקירה על ארכיטקטורות מבוססת SSM (הייתי זאת לפחות ב 3 הסקירות הקודמות). כמו שאתם יודעים הטרנספורמרים השתלטו היום גם על תחום הויז'ן והחלק הארי של מודלי SOTA בתחום מבוסס על הטרנספורמרים. הטרנספורמרים החליפו את רשתות קונבולוציה(CNN) ששלטו בתחום הויז'ן עד 2020 בערך. למרות שיש טוענים שבכל מודל ויז'ן עובד יש או איזה backbone מבוסס CNNs או שמכיל רכיב כמו attention לוקאלי (דרך להתחשב ב-inductive bias שיש בדאטה ויזואלי שאותה מנצלים CNNs), עדיין השליטה של הטרנספורמרים בויז'ן נראית די מוחלטת.

אוקיי, אתה זוכרים שהמטרה של ממבה שהמודלים שקדמו לה היה הדואליות(ראו הסבר מפורט בסקירות הקודמות) המאפשרת אימון מקבילי לצד היסק (inference) מהיר. בגדול מחליפים את מנגנון ה-attention שיש בטרנספורמרים במודל מבוסס SSM שמאפשר דחיסה של כל הזכרון(עד טוקן הנחזה) בוקטור אחד ובכך מאפשרים היסק מהיר (וגם אימון מקבילי).

אז למה לא נעשה את אותו הדבר עבור הדאטה הויזואלי? זה בדיוק מה שהמאמר מנסה לעשות. למעשה המאמר משלב CNNs (הרי נפטרים מהטרנספומרים) עם מנגנון דחיסת הזכרון המבוסס SSM. מה שקצת משעשע שבמהלך האימון מודלי בסגנון ממבה מופעלים דרך קונבולוציה ארוכה (הוסבר בסקירות הקודמות בהרחבה) אז קיבלנו בסוף רשת קונבולוציה טהורה (לפחות באימון).

מה הבעיה העיקרית עם הכנסה של קצת -SSM -ים למודלים ויזואליים? הכיוון!! הרי עבור שפה טבעית וגם עבור אודיו די ברור שעבור טוקן נתון צריך ״לזכור״ את הטוקנים מתחילת הטקסט/אודיו. בתמונה לטוקן נתון (פאץ') ניתן לבנות זיכרון מכיוונים שונים (הסדר חשוב בממבה וב-SSM-ים אחרים). אפשר להתחיל מלמעלה, או מלמטה שלה תמונה, ללכת ימינה או שמאלה. לא ברור מה הכי טוב מבחינת ביצועים ואז VMamba משלב אותם. לפאץ' נתון מתחילים מפאץ' השמאלי והעליון והפאצ'ים נכנסים ל-SSM בכיוון ימין-מטה. משלבים את ה-SSM הזה עם ה-SSM שנבנה החל מהפאץ' התחתון מימין כאשר כאשר הפאצ'ים נכנסים ל-SSM מכיוון שמאל-למעלה. ככה בונים את החלק ה-SSM של VMamba. מעניין למרות שהמאמר נקרא ממבה הוא טוען שהוא משתמש בארכיטקטורה של S6 שקדמה לממבה (ההבדל הוא תלות של מטריצות B ו- C של SSM בייצוג הטוקן הנוכחי).

שאר הדברים הדי סטנדרטיים: חלוקת תמונה לפאצ'ים, הפעלה כמה סבבי downsampling המבוצע עם בלוקים המכילים קונבולוציות 3x3, SSM, חיבור resnet וכמה שכבות לינאריות.

זהו זה, סוף הסקירה מקווה שנהנתם לקרוא את סדרת ממבה…