# Review 195: Can Mamba Learn How to Learn? A Comparative Study on In-Context Learning Tasks

**Paper: https://arxiv.org/abs/2402.04248v2**

https://arxiv.org/abs/2402.04248

אוקיי, סוקרים מאמר הבא בסדרת ממבה (מה שבא אחרי). בניגוד להצהרותיי בסוף סקירתי הקודמת לא תהיי Mamba MoE אלא מאמר אחר. הסיבה היא שלדעתי כמות המאמרים על MoE היא גדולה מאוד והמאמר הזה רק מציע להלביש אותו על Mamba ללא חידושים מעניינים אחרים אז החלטתי לדלג.

המאמר שנסקור היום בודק את האם מודלים המבוססים על ארכיטקטורת ממבה על למידת in-context (או ICL). למעשה ICL היא יכולת של מודל לבצע למידה על בסיס כמה דוגמאות בודדות (גם נקרא למידת few-shot) ללא שינוי של משקלי המודל. בגדול יכולת זו של הטרנספורמרים לא מאוד מפתיעה כי ה״חיזויים״ שלהם תלויים ביחסים בין חלקי הדאטה השונים(טוקנים) באופן מפורש באמצעות מנגנון ה-attention שלהם. כמובן יש מחקרים לא מעטים ומעניינים שחוקרים את התופעה המרתקת הזו ואני ממליץ לכם בחום להעיף מבט.

לעומת זאת הארכיטקטורה של ממבה לא לוקחת את היחסים בין הטוקנים השונים של הדאטה באופן מפורש ודוחסת את ה״עבר״ בוקטור אחד אז היכולת שלה לבצע ICL היא פחות אינטואיטיבי. זה אכן פחות קורה. המאמר בדק כמה ארכיטקטורות מבוססות (SSM (state-space models כמו ממבה, S4 וגם S4-mamba ללא מנגנון attention של הטרנספורמרים והשוו את יכולות ICL שלהם עם ארכיטקטורות הברידיות: כלומר שילוב של ממבה עם מנגנון ה-attention של הטרנספורמרים.

איך משלבים ממבה עם הטרנספורמר? המאמר בדק שתי גישות (די דומות). בגישה הראשונה הוא החליף את MLP שיש בבלוקי טרנספורמר אחרי ה-attention במנגנון של ממבה. הגישה השנייה (הכי מוצלחת) הנקראת MambaFormer בנוסף מחליפה את הקידוד המיקומי (positional encoding) של עוברים הטוקנים בממבה נוספת.

כאמור MambaFormer הגיע לביצועים הטובים ביותר מכל הארכיטקטורות הלא היברידיות (הטרנספורמר הטהור וכמה וריאנטים של SSM) באופן לא מפתיע בכלל. הרי MLP (רק 2 שכבות) ממדלים הפעולה די פשוטה ו-mamba היא למעשה מנגנון של זכרון הדוחס את המידע המהותי (בתקווה) של העבר (בטוקנים הקודמים). לא פלא שזה ניצח את כולם.

נשאר לנו רק לציין איזה משימות ניתנו למודלים אלו כדי לבחון את יכולות ICL שלהם. אחת המשימות היא לתת למודל כמה זוגות של (x, f(x)) עבור פונקציה f לינארית ולבקש ממנו לחשב (f(x עבור x-ים נוספים. משימה אחרת היתה לתת לה נקודות שנדגמו מ Gaussian Mixture מסוים ולבקש ממנו לדגום עוד נקודות. טבלה עם כל המשימות מצורפת לפוסט.

נתראה בסקירה ממבה הבאה והאחרונה (לא בחרתי עדיין).