# Review 206: SSAMBA: SELF-SUPERVISED AUDIO REPRESENTATION LEARNING WITH MAMBA STATE SPACE MODEL

**Paper: https://arxiv.org/abs/2405.11831v2**

https://arxiv.org/abs/2405.11831

המאמר הזה משך את תשומת ליבנו כי שמו דומה לממבה, ארכיטקטורה מעניינת שפרצה לתודעתנו לפני כחצי שנה וכבר יצאו עשרות מאמרים המשלבים אותה עבור מגוון דומיינים ומגוון משימות. והפעם התחום הוא אודיו והמחברים משתמשים בארכיטקטורת ממבה למטרת בניית ייצוג חזק של אות אודיו.

השאלה הראשונה שצריך לשאול כאן - מה הוא ייצוג חזק של דאטה. בהקשר זה באופן די טבעי ייצוג חזק של דאטה מקודד את התכונות החשובות שיש בדאטה כלומר דוחס את המידע המהותי שיש בדאטה בצורה יעילה. ייצוג זה נבנה על ידי מודל (מבוסס ממבה כאמור) ויכול לשמש אותנו לאימון של משימות נוספת על אותות אודיו. כלומר במקום לאמן מודל למשימה מסוימת על דאטה עצמו נאמן אותו על הייצוג הלטנטי של הדאטה (אמבדינג). דרך אגב התחום בלמידה מכונה העוסק בבנייה של ייצוגים אלו נקרא למידת הייצוג או representation learning.

כמו שאתם בטח זוכרים ממבה אמור לקבל כקלט אמבדינגים של טוקנים. בשפה טבעית כל טוקן הוא תת-מילה או מילה מוגדרים על ידי המילון, עבור תמונה הטוקנים הם פאצ'ים של תמונה (בסדר מסוים) אבל מה אנו עושים עם אות האודיו? האמת משהו די סטנדרטי - מחלקים את האות שלנו למקטעים זרים שכל קטע הוא כמה שניות. לאחר מכן מעבירים כל מקטע כזה דרך התמרת פורייה ולאחר מכן דרך טרנספורמציית מל (Mel transform). בגדול טרנספורמציית מל מדגישה את התדרים שהאוזן האנושית מסוגלת לשמוע. לאחר מכן מעביר את התוצאה של מל דרך שכבה לינארית ומוסיפים קידוד מיקומי (positional encoding) המקודד מיקומו של כל טוקן אודיו בסדרה.

לאחר מכן מעבירים את התוצאה דרך שכבת ממבה (די סטנדרטית - ניתן למצוא את תיאורה בהרבה מקומות כולל בסקירותיי (לינק) הרבות בנושא זה). בדומה למודל ממבה לראייה ממוחשבת (שם המצב אפילו יותר מורכב כי הפאצ'ים של תמונה הם דו-מימדיים) כאן מכניסים את ייצוגי הטוקנים לממבה בשני ״סדרים״: מהתחלה עד הסוף (forward) ומהסוף להתחלה (backward) ומשלבים אותם כדי לבנות את הפלט.

מה שיוצא אחרי כמה שכבות של ממבה הוא למעשה ייצוג תלוי הקשר (contextualized) של הטוקן וכאמור ניתן לנצל אותו לאימון מודלים למגוון משימות ייעודיות.

אבל איך מאמנים את המודל המפיק את הייצוג הזה. בצורה די סטנדרטית האמת. ממסכים חלק מהטוקנים (כמו באימון של מודלי שפה) ואז בונים לוס המורכב משני חלקים:

הלוס הניגודי (contrastive loss): כאן המטרה לקרב את הייצוג של הטוקן הממוסך לייצוגו (מהאיטרציה הקודמת של אימון) ובאותו הזמן להרחיק אותו מהייצוגים של הטוקנים האחרים. ניתן להשיג את היעד הזה עם פונקציית לוס, לראשונה הוצגה במאמר InfoNCE (לינק) לפני 8 שנים בערך.

כאן מנסים לקרב את ייצוג הטוקנים הממוסכים עם ייצוגו (מהאיטרציה אימון הקודמת). המרחק בין חיזוי הייצוג והייצוג עצמו מוגדר כ L2 כלומר אוקלידי.