# Review 191: Hyena Hierarchy: Towards Larger Convolutional Language Models

**Paper: https://arxiv.org/abs/2302.10866v3**

https://arxiv.org/abs/2302.10866

היום סוקרים את המאמר השביעי בסדרה וכאן אני חייב להודות שלקח לי הרבה מאוד זמן לצלול למאמר הזה לעומק למרות שטכנית המאמר לא מורכב במיוחד (בטח לא קרוב ל Hippo). אבל המאמר כתוב בצורה נוראית: מצד אחד הוא עמוס בפרטים לא מהותיים ומצד שני נעשה מאמץ ניכר (על ידי המחברים) להסתיר את הפרטים המהותיים עם מלל אינסופי. לא יודע האם זה נעשה בזדון או לא אבל המאמר הזה לקח לי בערך פי 4 יותר זמן ממאמר ממוצע שזה הרבה סטיות תקן מהממוצע (יש לי מדגם די גדול).

אחרי ששחררתי את הקיטור אפשר להתחיל לסקור את המאמר זה שמציע הכללה חמודה ל H3 שסקרנו קודם. H3 היה די נחמד אבל עדיין הביצועים שלו לא היו בשמיים עבור כמה משימות על הדאטה בעלי אורך הקשר ארוך מאוד. אז באו לנו מחברי Hyena והציעו לשפר את ביצועי H3 אך לא במחיר של עלייה ניכרת במשאבי חישוב והזיכרון.

אוקיי, אז מה הם הציעו בעצם? אתם זוכרים שב-H3 אנו לקחנו וקטורי מפתח עבור הטוקנים בתוך חלון ההקשר (=מטריצה K) העברנו אותם דרך SSM (State-Space Models) ואז הכפלנו אותם בווקטורי שאילתה (=מטריצה Q) והעברנו את התוצאה דרך SSM נוסף עם מטריצה A אחרת ואת התוצאה הכפלנו בווקטורי ערך עבור כל הטוקנים בתוך חלון ההקשר (=מטריצה V)? כל המנגנון הזה הוא למעשה attention לינארי.

אז ההכללה הראשונה המוצעת במאמר היא הגדלת מספר הוקטורים שעליהם מופעלת SSM (בצורה לא מפורשת - נדבר על זה עוד מעט) ל N. כלומר יש לנו 1+N הטלות של ייצוג הטוקנים (אחת עבור מטריצת הערך V). אחרי שיש לנו את ההטלות האלו מפעילים עליהם מה שבמאמר נקרא Short Convolution (קונבולוציה קצרה) בציר הטוקנים. זה נעשה כנראה כדי ללמוד את האינטראקציות בין הטוקנים הסמוכים (המאמר לא מסביר כלום לגבי זה).

מפה העניינים קצת מסתבכים. אנו לוקחים מטריצת הערך V מההטלה האחרונה ומפעילים עליהם SSM (אותה מערכת דינמית לינארית) אבל בצורה לא מפורשת. מה זה אומר אבל? אנו יודעים שהפעלת SSM לסדרה של L טוקנים שקולה להפעלה של קרנל קונבולוציה באורך L על ייצוגי טוקנים אלו. קרנל קונבולוציה זה מוגדר על ידי המטריצות המגדירות את ה-SSM (שזה A, B, C). אז ניתן להגדיר SSM בצורה לא מפורשת דרך הקרנל הזה. צריך לזכור פעולה זו שקולה להכפלת וקטורים, המרכיבים מטריצת ערך V, במטריצת קונבולוציה גדולה (= שזה אותו מנגנון של attention לינארי).

למשל ב-H3 (שסקרנו בפעם הקודמת) היו לנו שני SSMs (עם מטריצה אלכסונית ועם מטריצת הזזה ב-1) ומתברר שניתן לייצג אותם בצורה לא מפורשת עם קרנל שהוא מכפלה של שתי מטריצות שכל אחת מהן היא מכפלה של מטריצה אלכסונית במטריצת Toeplitz. מה שמיוחד במטריצת Toeplitz היא שכל שורה בה כי הזזה שמאלה של השורה הקודמת. תכונה מעניינת של כל מטריצה Toeplitz היא שהיא מהווה ייצוג של קרנל קונבולוציה.

אז המחברים לקחו את הייצוג הלא מפורש של SSM ובנו אותו מ- N מכפלות של מטריצות אלכסוניות ומטריצות Toeplitz (שונות). כלומר מתחילים מטריצה V עבור הטוקנים מפעילים עליה מיפוי $H$ לינארי (= קרנל קונבולוציה) די מסורבל. כלומר H הוא הרכבה של N מיפויים $H\_i, i=1,...N$ לינאריים שכל אחת מהן היא קונבולוציה המיוצגת על ידי מטריצה Toeplitz (מס' i) ומכפלת התוצאה איבר-איבר בהטלה מספר i של וקטורי הטוקנים. במאמר כל הסיפור הזה נקרא Hyena operator מסדר N.

אוקיי, מה הבעיה העיקרית עם הגישה הזה? זה דורש הרבה זיכרון בטח עבור N גדול יחסית. אז המאמר מציע פתרון מאוד אלגנטי. במקום ללמוד את כל N קרנלים אלו בצורה מפורשת נגדיר אותם באמצעות רשת נוירונים רדודה (fully-connected). גם נוכל לשלוט על מספר פרמטרים וכך לשמור על זיכרון קבוע פחות או יותר לכל ערך של N. כך ניצור את כל N קרנלים עם רשת אחת בלבד. ארכיטקטורת רשת רדודה זאת היא די מיוחדת והיא מכילה פונקציות אקטיבציה מחזוריות (כדי ליצור קונבולוציות עם תדרים גבוהים).

בנוסף מכפילים קרנל זה (איבר איבר) בפונקציה מעריכית עם פרמטר חיובי דלטה $\exp(-\delta t)$ בציר הטוקנים. הכפלה זו באה לשקף דעיכה בהתחבשות בטוקנים(=attention( ככל המרחק בינם לבין הטוקן החזוי יורד. המאמר משתמש בכמה אופרטורי Hyena (ערוצים) במקביל עם מקדמי $delta$ שונים המבטאים קצבי דעיכה שונים של attention. כל אופרטור כזה מופעל על וקטורי קידוד מיקומי (positional encoding).

ודבר אחרון: כל הקונבולוציות מחושבות דרך FFT(Fast Fourier Transform) וגם IFFT כמו במאמר של H3 (כי זה פשוט יותר מהיר). כמובן כל SSM (גם לא מפורש) מופעל על כל מימד של ייצוג הטוקנים שטיפה מסבך את התיאור אבל עדיין הכל נשאר לינארי.