# Review 45: Rethinking Attention With Performers

**Paper: https://arxiv.org/abs/2009.14794v4**

פינת הסוקר:

המלצת קריאה ממייק: חובה לחובבי הטרנספורמרים ולאנשי NLP.

בהירות כתיבה: גבוהה.

רמת היכרות עם כלים מתמטיים וטכניקות של ML/DL הנדרשים להבנת מאמר: נדרשת היכרות בסיסית עם תורת הקרנלים, הבנה טובה בפעולת ליבה בטרנספורמרים (self-attention).

יישומים פרקטיים אפשריים: ניתן להשתמש בגישה המוצעת במאמר עבור כל משימה בה הסיבוכיות הריבועית של מנגנון self-attention של הטרנספורמר הינה בעיה מבחינת משאבי חישוב.

פרטי מאמר:

פורסם בתאריך: 09.03.21, בארקיב.

הוצג בכנס: ICLR 2021.

תחומי מאמר:

טרנספורמרים בעלי סיבוכיות חישובית נמוכה.

כלים מתמטיים, מושגים וסימונים:

מנגנון SA - self-attention.

קרנלי סופטמקס (softmax kernels).

פיצ'רים חיוביים אורתוגונליים רנדומליים (Positive Orthogonal Random Features).

מבוא ותמצית מאמר:

טרנספורמר הינו ארכיטקטורה של רשתות נוירונים עמוקות שהוצעה בשלהי 2017 במאמר "Attention is what you need". מאז הטרנספורמים כבשו את עולם ה-NLP והפכו לארכיטקטורה כמעט דפולטית בתחום. רוב המוחלט של מאמרי NLP של השנים האחרונות משתמשים בטרנספורמרים בצורה זו או אחרת. לאחרונה הטרנספורמרים התחילו לפלס את דרכם גם לדומיין הויזואלי והופיעו בכמה מאמרים שחלקם סקרתי (Image is Worth 16x16 Words, TransGAN, Image Processing Transformer).

הקלט לטרנספורמר הינו סט או סדרה של עצמים (מילה, תת-מילה, פאטץ' בתמונה, דגימות אודיו וכו') שכל אחד מהם מיוצג על ידי וקטור. הלב של הטרנספורמר הינו מנגנון self-attention שמטרתו כימות קשרים בין איברים שונים בסט ובסדרה. המטרה של הטרנספורמר הינה הפקה של ייצוג וקטורי של כל איבר בסדרה/סט, התלוי באיברים האחרים (מה שנקרא contextualized embedding ב-NLP). דרך אגב לאחרונה יצא מאמר, שהראה שהכוח של מנגנון self-attention נובע משילובו עם skip-connections ושכבות fully-connected. בנוסף נציין כי כאשר הקלט הינו בעל סדר אינהרנטי בין איבריו (כמו טקסט או תמונה), אז מוסיפים לוקטור הייצוג של כל איבר גם וקטור המכיל מידע על מיקומו בסדרה (Positional encoding - PE). כאשר הקלט הינו סט ללא חשיבות לסדר (אינווריאנטי לתמורות), לא נדרש PE.

מאחר ובשלב הראשון מנגנון SA מחשב את הדמיון של כל איבר בסדרה לכל איבר אחר בסדרה, הסיבוכיות של שלב זה הינה ריבועית במונחי אורך הסדרה (נסמן את אורך הסדרה ב-L). סיבוכיות זו עלולה להיות בעייתית עבור סדרות ארוכות מבחינת משאבי חישוב וזכרון הנדרשים. בעיה זו מחריפה עבור ארכיטקטורות המורכבות ממספר שכבות של טרנספורמרים. אגב, סוגיה זו מהווה את אחד המכשולים המהותיים (בנוסף לכך שהטרנספורמר, בצורתו הקלאסית, לא בנוי לניצול קשרים לוקאליים הקיימים בתמונות, אך זה ניתן לטיפול על ידי אימוץ שיטות אימון מתוחכמות) המונעים את השתלטות הטרנספורמרים גם על הדומיין הויזואלי. הסיבה לכך טמונה במספר הפאטצ'ים (איברים בסדרה) הגבוה בתמונה ברזולוציה גבוהה - המימוש הסטנדרטי של מנגנון SA עלול להיות כבד מאוד הן מבחינה חישובית והן מבחינת הזכרון הנדרש).

בשנה האחרונה יצאו מספר מאמרים שהציעו וריאנטים זולים יותר חישובית של הטרנספורמר כמו Linformer ו-Reformer. כדי להוריד את הסיבוכיות הריבועית של הטרנספורמר, רוב המאמרים הניחו הנחות על תכונות של הקשרים בין האיברי הסדרה או/ו על מטריצות Q, K ו-V המשתתפות בחישוב של SA. לטענת מחברי המאמר הנסקר כל הוריאנטים "קלים חישובית" של הטרנספורמר, שנבדקו על ידיהם, הפגינו ביצועים ירודים משמעתית יחסית לגרסתו המקורית (היקרה חישובית) של הטרנספורמר. המאמר טוען שהסיבה לביצועים חלשים אלו הינה אי-קיום של התנאים עליהם מתבססים וראינטים אלו.

כותבי המאמר אינם מניחים שום הנחה על תכונות/מבנה של הקשרים בין איברים ומציעים מסגרת מתמטית ריגורוזית למציאת קירוב למטריצת attention (המחושבת על ידי מנגנון SA) בסיבוכיות לינארית במונחי אורך הקלט. בנוסף, ניתן לשחק עם הפרמטרים של קירוב זה ולהגיע לכל דיוק רצוי בשערוך של מטריצת attention. יתרה מזו, המאמר מוכיח כי שקירוב זה הינו:

אומדן בלתי מוטה (או ממש קרוב לזה) למטריצת attention.

מתכנס בצורה יוניפורמית (אותה מהירות התכנסות לכל איברי מטריצת attention ולכל טווח ערכי ה-attention).

בעל שונות נמוכה.

הסבר של רעיונות בסיסיים:

כאמור בשלב הראשון של חישוב מטריצת attention, פעולת softmax מחושבת על מכפלת המטריצות \*Q ו- \*K (משוחלפת). מטריצות \*Q ו- \*K מורכבות מהמכפלות של מטריצות Query ומטריצת Key (המסומנות על ידי Q ו-K בהתאמה) על וקטורי הייצוג של הקלט q\_i ו-k\_j. למעשה כל המכפלות הפנימיות מנורמלות ב- d1/2 אך זה לא משנה את עיקרי החישוב. כלומר פעולת softmax מופעלת על המטריצה (נסמן אותה כ-A), שאיבר {ij} שלה הינו מכפלה פנימית של וקטורי q\_i ו-k\_j. נציין שגודל מטריצה זו היא LxL, כאשר L הינו אורך הקלט. לאחר מכן, התוצאה של פעולת מטריצה A מוכפלת במטריצה \*V שבנויה ממכפלות של וקטורי ייצוגי האיברים במטריצת V (מטריצת Value). הגודל של מטריצת \*V הינו L×d, כאשר d הינו מימד של וקטורי הייצוג. ניתן לראות כי סיבוכיות זמן וגודל זכרון הנדרש הם (O(L^2 . וזה לב הבעיה עם הטרנספורמטורים עבור קלט ארוך כמו פסקה שלמה של טקסט או כל הפאטצ'ים של תמונה ברזולוציה גבוהה. המאמר מציע שיטה לקרב את החישוב של softmax על המכפלה של \*Q ו-\*K משוחלפת על ידי מכפלה של שתי מטריצות 'Q ו-'K בגודל של Lxr, כאשר r הרבה יותר קטן מ-L. זה מאפשר להחליף את סדר המכפלה של המטריצות בחישוב SA:

מכפילים מטריצה V בגודל Lxd במטריצה 'K משוחלפת בגודל rxL. כתוצאה מכך מקבלים מטריצה 'A בגודל rxd.

מכפילים את המטריצה 'A שנתקבלה במטריצה 'Q בגודל rxL.

קל לראות שהסיבוכיות של הזכרון ושל החישוב במקרה זה אינה לינארית ב-L (כאשר r<<L).

אבל השאלה המהותית כאן היא: איך ניתן לבנות מטריצות 'Q ו-'K כך שמכפלתן תהווה קירוב בעל תכונות המוזכרות לעיל (בלתי מוטה, בעל קצב התכנסות יוניפורמית שונות קטנה). מחברי המאמר מציעים שיטה, הנקראת ++FAVOR, לקירוב של מטריצה A, שאיבריה הם ערכי ה-softmax כאשר הארגומנטים שלו הם המכפלות הפנימיות של וקטורי q ו- k. למעשה, המאמר מציע שיטה יותר כללית לקירוב של כל פונקציה מהצורה (K(q, k, כאשר K זה קרנל (פונקציית בעלת תכונות מסוימות) חיובי. הקירוב למעשה מהווה תוחלת של מכפלה פנימית של (φ(q ו- (φ(k (מסומנת E(q, k) כאשר φ הינה פונקציה אקראית (randomized) מ- Rd ל- R. ד״א זה די מזכיר ייצוג קרנל באמצעות Random Fourier Features למי שמכיר. המאמר מציע לקחת את פונקצית מהצורה הבאה:

(1)

כאשר

fi, i=1,..l הינן פונקציות R→R.

h הינה פונקציה Rd→R.

ϖi, i=1..., m - הינם וקטורים, המוגרלים (פעם אחת לאורך כל החישוב) מהתפלגות D על Rd. ברוב המקרים   
התפלגות D הינה איזוטרופית, כלומר פונקצית ההתפלגות שלה קבועה על ספרה (sphere).

לדוגמא, אם נקח ()h≡1, f1=cos(), f2=sin, ו-D הינה התפלגות גאוסית סטנדרטית, אז נקבל קירוב של מה שנקרא קרנל גאוסי Kgauss. במקרה שלנו אנו צריכים למצוא קירוב ל- (SM(x, y) = exp(xTy (עד כדי הנרמול). אם נשים לב כי

SM(x, y) = exp(||x||2/2) Kgauss(x, y) exp(||y||2/2) (2)

אז קל להראות כי (SM(x, y ניתן לקרוב על ידי פונקציה, המוגדרת על הפונקציות הבאות באמצעות הנוסחה (1):

h(x) = exp(||x||2/2), f1=cos(), f2=sin() (3)

אז למעשה הצלחנו לקרב את איברי מטריצות \*Q ו-\*K משוחלפת על ידי מכפלה פנימית של וקטורים, המחושבים מוקטורי qi ו- vj (עם פונקציית phi). נוכל אז לבצע את מכפלת המטריצות בביטוי של מטריצת attention בסדר אחר, ובכך הורדנו את הסיבוכיות ללינארית במונחי אורך הקלט. אבל יש פה קאטץ' קטן: softmax למעשה יותר צירוף לינארי קמור (שכל מקדמיו חיוביים ומנורמלים) של המכפלה של \*Q ו-\*K משוחלפת. כאשר אנו מחליפים את החישוב הזה על ידי הקירוב שיכול לקבל כל ערך (גם שלילי), זה עלול להיות בעייתי ולגרום לאי דיוקים רציניים במיוחד במקומות ש ערך ה- softmax קרוב לאפס. אם ניזכר ש softmax מודד דמיון בין וקטורי query לוקטורי key בין איברים שונים, אזי סביר להניח שרוב ערכיו יהיו קרובים לאפס. המאמר גם מראה שאם משתמשים בקירוב (3) אז אי הדיוקים של הקירוב, יחסית לערכים האמיתיים של softmax, הינם די משמעותיים.

כלומר לא רק שאנחנו צריכים לקרב את החישוב של softmax אלא לעשות זאת באמצעות פונקציות לא שליליות. המאמר מציע להשתמש בקירוב הבא:

שניתן על ידי

המאמר מראה שקירוב ה-softmax דרך הביטוי, הניתן על ידי שתי המשוואות האחרונות, מצליח לקרב את הערכים האמיתיים של מטריצת ה-attention בצורה יוניפורמית ועם שונות נמוכה. כדי לגרום לקירוב להיות יותר מדויק בהינתן אותו מספר של וקטורים המוגרלים מהתפלגות גאוסית סטנדרטית ϖi, i=1..., m (פעם אחת בלבד לאורך כל הדרך), המאמר מציע לבצע תהליך אורתוגונליזציה של וקטורים אלו. אחד הדרכים לעשות זאת היא להשתמש בשיטת גרם-שמידט.

לבסוף, המאמר מוכיח בצורה ריגורוזית (באמצעות כלים די לא טריוויאליים את התכונות התיאורטיות ״הטובות״ של הקירוב הזה (רוב המאמר זה הוכחות - בערך 30 עמודים).

הישגי מאמר:

המאמר הראשון (למיטב ידיעתי) שהצליח להקטין את סיבוכיות החישוב (והאכסון) של מטריצת ה-attention בטרנספורמר ללינארית במונחי אורך סדרת הקלט ללא הנחות כלשהן על מטריצות Key, Query, Value ועל ערכי attention עצמם.

נ.ב.

מאמר מציע שיטה להקטין את סיבוכיות של הטרנספורמר ללינארית ומוכיח את כל טענותיו גם (!!) בצורה רגורוזית. המאמר לא פשוט לקריאה אך לשמחתנו כדי להבין את העיקר לא צריך להתעמק בפרטי ההוכחות (5 -6 העמודים הראשונים מספיקים).