סקירה זו היא חלק מפינה קבועה בה אני סוקר מאמרים חשובים בתחום ה-ML/DL, וכותב גרסה פשוטה וברורה יותר שלהם בעברית. במידה ותרצו לקרוא את המאמרים הנוספים שסיכמתי, אתם מוזמנים לבדוק את העמוד שמרכז אותם תחת השם deepnightlearners.

לילה טוב חברים, היום אנחנו שוב בפינתנו deepnightlearners עם סקירה של מאמר בתחום הלמידה העמוקה. היום בחרתי לסקירה את המאמר שנקרא:

RETHINKING ATTENTION WITH PERFORMERS

פינת הסוקר:

המלצת קריאה ממייק: חובה לאוהבי הטרנספורמרים.

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:

לינק למאמר: זמין להורדה.

לינק לקוד: זמין כאן.

פורסם בתאריך: 09.03.21, בארקיב.

הוצג בכנס: ICLR 2021.

תחומי מאמר:

- טרנספורמרים בעלי סיבוכיות חישובית נמוכה.

כלים מתמטיים, מושגים וסימונים:

- מנגנון self-attention - SA.

- 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, TransGAN, Image_is_Worth_16x16_Words 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 ייצוג של כל איבר, וקטור המכיל מידע על מיקומו בסדרה (Positional encoding - PE). כאשר הקלט הינו סט ללא חשיבות לסדר (אינווריאנטי לתמורות), PE לא נדרש.

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. דרך אגב סוגיה זו מהווה 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 גם חישובית וגם מבחינת הזכרון הנדרש).

בשנה האחרונה יצאו כמה 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 המשתתפים 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] (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 הטווח של ערכי ה-attention).
- בעל שונות נמוכה.

הסבר של רעיונות בסיסיים:

כאמור בשלב הראשון של חישוב מטריצת attention, פעולת softmax מחושבת על המכפלת של מטריצת ***Q** ו- ***K** (משוחלפת). מטריצות ***Q** ו- ***K** מורכבות המכפלות 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- $d^{1/2}$ אך זה לא משנה את

עיקרי החישוב. כלומר פעולת softmax מופעלת מטריצה (נסמן אותה כ-**A**). שאיבר {ij} שלה הינו מכפלה פנימית של וקטורי q_i ו-k_j. נציין שהגודל של 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 של המכפלה של ***Q** ו-***K** משוכלפת 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:

1. מכפילים מטריצה **V** בגודל Lxd במטריצה **K'** משוכלפת בגוגל rxL. כתוצאה מכך מקבלים מטריצה **A'** בגודל rxd.
2. מכפילים מטריצה **A'** במטריצה **Q'** בגודל rxL

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

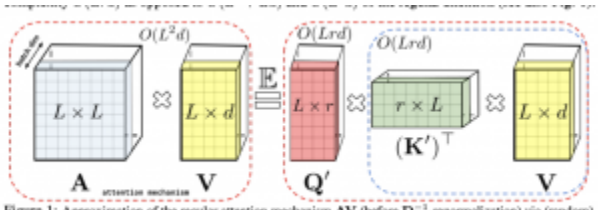

Figure 1: Approximation of the regular attention mechanism AV (before D^{-1}-renormalization) via (random) feature maps. Dashed-blocks indicate order of computation with corresponding time complexities attached.

אבל השאלה המהותית כאן היא: איך ניתן לבנות מטריצות **Q'** ו-**K'** כדי שמכפלתן תהווה קירוב בעל תכונות המוזכרות לעיל (בלתי מוטה, בעל קצב התכנסות יוניפורמית שונות קטנה). המחברים 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- $\mathbf{R}^d$ ל- **R**. ד"א זה די מזכיר ייצוג קרנל באמצעות Random Fourier Features למי שמכיר. המאמר מציע לקחת את פונקצית מהצורה הבאה:

$$\phi(\mathbf{x}) = \frac{h(\mathbf{x})}{\sqrt{m}}(f_1(\omega_1^\top \mathbf{x}), ..., f_1(\omega_m^\top \mathbf{x}), ..., f_l(\omega_1^\top \mathbf{x}), ..., f_l(\omega_m^\top \mathbf{x})) \qquad (1)$$

כאשר

- f_i, i=1,..l הינן פונקציות **R→R**.
- h הינה פונקציה $\mathbf{R}^d$**→R**.
- ϖ_i, i=1..., m - הינם וקטורים, המוגרלים (פעם אחת לאורך כל החישוב) מהתפלגות D על $\mathbf{R}^d$.ברוב המקרים.

התפלגות D הינה איזוטרופית כלומר פונקצית התפלגות שלה קבועה על ספרה (sphere). לדוגמא אם ניקח h≡1,() f_1=cos(), f_2=sin, ו-D הינה התפלגות גאוסית סטנדרטית אז נקבל קירוב של מה שנקרא קרנל גאוסי K_{gauss}. במקרה שלנו אנו צריכים למצוא קירוב ל- SM(**x, y**) = exp($\mathbf{x}^T\mathbf{y}$) (עד כדי הנרמול). עם נשים לב כי

$$SM(\mathbf{x}, \mathbf{y}) = \exp(||\mathbf{x}||^2/2)\, K_{gauss}(\mathbf{x}, \mathbf{y})\, \exp(||\mathbf{y}||^2/2) \qquad (2)$$

אז קל להראות כי SM(**x, y**) ניתן לקרב על ידי פונקצייה, המוגדרת על הפונקציות הבאות באמצעות נוסחה (1):

$$h(x) = \exp(||\mathbf{x}||^2/2),\ f_1=\cos(),\ f_2=\sin() \qquad (3)$$

אז למעשה הצלחנו לקרב את איברי מטריצות **Q*** ו-**K*** משוכלפת על ידי מכפלה פנימית של וקטורים, המחושבים מוקטורי qi ו- vj (עם פונקציית phi). אז נוכל לבצע את מכפלת המטריצות בביטוי של מטריצת attention בסדר אחר ובכך הורדנו את הסיבוכיות ללינארית במונחי אורך הקלט. אבל יש קאטץ' קטן כאן: softmax למעשה יותר צירוף לינארי קמור (שכל מקדמיו חיוביים ומנורמלים) של המכפלה של **Q*** ו-**K*** משוכלפת. כאשר אנו מחליפים את החישוב הזה על ידי הקירוב שיכול לקבל כל ערך (גם שלילי). זה עלול להיות בעייתי ולגרום לא אי דיוקים רציניים במיוחד במקומות ש ערך ה- softmax קרוב לאפס. ואם ניזכר של softmax מודד דמיון בין וקטורי query לוקטורי key בין איברים שונים, 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 אני לא רק צריכים לקרב את החישוב של softmax אלא לעשות זאת באמצעות פונקציות לא שליליות. המאמר מציע להשתמש בקירוב הבא:

$$\mathrm{SM}(\mathbf{x},\mathbf{y}) = \mathbb{E}_{\omega\sim\mathcal{N}(0,\mathbf{I}_d)}\Big[\exp\Big(\omega^\top\mathbf{x}-\frac{\|\mathbf{x}\|^2}{2}\Big)\exp\Big(\omega^\top\mathbf{y}-\frac{\|\mathbf{y}\|^2}{2}\Big)\Big]$$

שניתן על ידי

$$h(\mathbf{x}) = \frac{1}{\sqrt{2}}\exp(-\frac{\|\mathbf{x}\|^2}{2}),\ l=2,\ f_1(u)=\exp(u),\ f_2(u)=\exp(-u)$$

המאמר מראה קירוב ה-softmax ביטוי הניתן על ידי שתי משוואות 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 ϖ_i, i=1..., m (פעם אחת בלבד לאורך כל הדרך), מאמר מציע לבצע תהליך של אורתוגונוליזציה של וקטורים אלו. אחד הדרכים לעשות זאת היא להשתמש בשיטת גרם-שמידט.

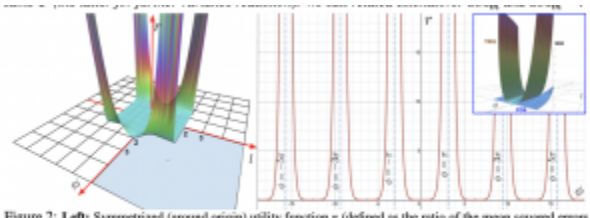

Figure 2: **Left:** Symmetrized (around origin) utility function r (defined as the ratio of the mean squared errors (MSEs) of estimators built on: trigonometric and positive random features) as a function of the angle ϕ (in radians) between input feature vectors and their lengths l. Larger values indicate regions of (ϕ, l)-space with better performance of positive random features. We see that for critical regions with ϕ large enough (small enough softmax-kernel values) our method is arbitrarily more accurate than trigonometric random features. Plot presented for domain $[-\pi,\pi]\times[-2,2]$. **Right:** The slice of function r for fixed $l=1$ and varying angle ϕ. **Right Upper Corner:** Comparison of the MSEs of both the estimators in a low softmax-kernel value region.

לבוסף המאמר מוכיח בצורה רגורוזית (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 ריגוריזית. המאמר לא פשוט לקריאה אך לשמחתנו כדי להבין את העיקר לא צריך להתעמק בפרטי ההוכחות (5 -6 העמודים הראשונים מספיקים).

#deepnightlearners

הפוסט נכתב על ידי מיכאל (מייק) ארליכסון, Michael Erlihson, PhD.

מיכאל עובד בחברת סייבר Salt Security בתור Principal Data Scientist. מיכאל חוקר ופועל בתחום הלמידה העמוקה, ולצד זאת מרצה ומנגיש את החומרים המדעיים לקהל הרחב.