**מנגנון attention לא צץ יש מאין: סקירה של המאמר 'מקורו של self-attention, סקירה 512**

**המאמר היומי של מייק: 19.09.25**

**THE ORIGIN OF SELF-ATTENTION: PAIRWISE AFFINITY MATRICES IN FEATURE SELECTION AND THE EMERGENCE OF SELF-ATTENTION**

המאמר של היום הוא פחות תגלית ויותר סוג של ״ארכאולוגיה אינטלקטואלית״. הוא טוען שמנגנון ה-Self-Attention, שבלב ארכיטקטורת הטרנספורמר, לא היה המצאה שהופיעה כרעם ביום בהיר בשנת 2017. במקום זאת, הוא מייצג את השכלול של עיקרון חישובי עמוק וכללי יותר שהתפתח במשך שנים: מינוף של pairwise affinity matrices כדי לווסת את זרימת המידע. המאמר ממסגר מחדש את מנגנון ה-attention כמקרה ספציפי, ונלמד של מסגרת רחבה יותר, כאשר אב קדמון רעיוני מרכזי שלו הוא שיטה לבחירת פ'יצרים (feature selection) משנת 2015, הידועה בשם בחירת תכונות אינסופית (Infinite Feature Selection או Inf-FS).

כדי להבין את טיעון הליבה של המאמר, יש להבין תחילה את ההיגיון של Inf-FS. דמיינו את פ'יצרים של הדאטה שלכם כחברים בפרלמנט. שיטות מסורתיות היו מעריכות את התרומה של כל תכונה בנפרד. Inf-FS, לעומת זאת, בונה גרף שלם (fully connected) שבו כל פיצ'ר מחובר לכל אחרים. המשקל של כל קשר, אפיניטי, מכמת את היחסים ביניהן, למשל קורלציה סטטיסטית או דמיון סמנטי. תהליך זה יוצר מטריצה צפופה של יחסים זוגיים, הלא היא מטריצת אפיניטי A.

ה"אינסוף" ב-Inf-FS נובע מהאופן שבו מחושבת חשיבותה של כל תכונה. היא אינה מבוססת רק על קשריה הישירים. הציון של תכונה נקבע על ידי צבירת ההשפעה המופצת דרך כל המסלולים האפשריים בכל אורך אפשרי על פני הגרף כולו. זוהי המקבילה הרעיונית של טור חזקות של מטריצה. האיבר הראשון בסדרה לוקח בחשבון קשר ישיר (one-hop). האיבר השני לוקח בחשבון מסלולים באורך שני צעדים, כיצד תכונה אחת קשורה לאחרת. האיבר השלישי מתייחס למסלולים באורך 3 צעדים, וכן הלאה, עד אינסוף. סיכום כל אלה נותן מדד גלובלי למידת המרכזיות וההשפעה של כל תכונה במערכת התכונות כולה. התוצר הסופי הוא דירוג סטטי של התכונות החיוניות ביותר למבנה הכולל של הנתונים.

בשלב זה, המאמר ממפה את המבנה הזה על מנגנון self-attention. שכבת Self-Attention בונה גם היא גרף קשיר לחלוטין, אך הצמתים בו הם טוקנים בסדרת קלט, ולא קבוצה סטטית של פיצ'רים. גם היא מחשבת מטריצת אפיניטי זוגית, מה שאנו מכירים כציוני attention לפני הנרמול, המכמתת את היחסים בין כל זוג טוקנים.

כאן, המחברים מזהים את הקשרים וההבחנות המכריעים:

1. אפיניטי נלמד ודינמי: בניגוד למטריצת אפיניטי של Inf-FS, שלעיתים קרובות מוגדרת ידנית או נגזרת סטטיסטית, המטריצה במנגנון Self-Attention נלמדת. חשוב מכך, היא דינמית: מטריצת אפיניטי חדשה מחושבת עבור כל קלט בנפרד, בכל שכבה, במהלך ה-forward pass. היחסים אינם קבועים; אלא תלויי-הקשר.
2. אגרגציה בצעד-יחיד (Single-Hop Aggregation): שכבת קשב-עצמי יחידה מבצעת אגרגציה בצעד-יחיד. לאחר חישוב האפיניטי ונרמולו באמצעות Softmax, היא יוצרת ייצוג חדש לכל טוקן על ידי חישוב ממוצע משוקלל של כל הטוקנים האחרים. מבחינה מבנית, זה אנלוגי להתחשבות רק ב*איבר הראשון* בטור החזקות של Inf-FS, כלומר, ביחסים הישירים והמיידיים.
3. רב-צעדיות (Multi-Hop) באמצעות סטאקינג של שכבות: כיצד טרנספורמר לוכד יחסים עמוקים ועקיפים יותר? באמצעות סטאקינג של שכבות. הפלט של שכבת Self-Attention אחת הופך לקלט של השכבה הבאה. זה מאפשר למידע להתפשט על פני הגרף באופן הדרגתי. טרנספורמר עם שתי שכבות יכול למדל יחסים של שני צעדים, מודל עם שלוש שכבות יכול ללכוד מסלולים באורך שלושה צעדים, וכן הלאה. באופן זה, מושגת התפשטות רב-צעדית הדומה לזו של Inf-FS, אך באמצעות עומק ארכיטקטוני ולא באמצעות פתרון אנליטי של סכימת טור בצעד אחד.

מיסגור מחדש זה מציב את הקשב העצמי כתת-קבוצה אלגנטית ויישומית של פרדיגמת Inf-FS. Inf-FS מספק את המבנה המתמטי הכללי לניקוד מבוסס-זיקה, בעוד Self-Attention הוא מימוש ספציפי שעבר אופטימיזציה ללמידת ייצוגים (representation learning) בתוך מערכות עמוקות וגזירות.

המאמר מבסס את טענתו על ידי התחקות אחר דפוס מבוסס-אפיניטי זה בעולם לֹמידת מכונה. רשתות נוירונים לא-מקומיות (Non-Local Neural Nets), כלומר לא ConvNets, בתחום הראייה הממוחשבת מחשבות במפורש אפיניות זוגיות בין כל המיקומים המרחביים במפת פיצ'רים כדי לתפוס תלויות ארוכות-טווח, תוך שימוש במנגנון self-attention כמקור השראה. בדומה, רשתות attention גרפיות (Graph Attention Networks או GATs) התקדמו מעבר לשימוש במטריצת הסמיכויות (adjacency matrix) הקבועה של הגרף, ובמקום זאת למדו משקלי attention לכל קשת, ובכך יצרו למעשה מטריצת אפיניטי נלמדת וספציפית למשימה כדי לשקלל את העברת המסרים בין הצמתים.

בסופו של דבר, "מקורו של attention" מספק טיעון משכנע לקיומה של רציפות רעיונית. גאונותו של הטרנספורמר לא הייתה בהמצאת אפיניטי הזוגית יש מאין, אלא בשילובה המופתי בארכיטקטורה דינמית, סקלבילית, ונלמדת מקצה-לקצה. על ידי התחקות אחר שושלת יוחסין זו, המאמר אינו מפחית מהשפעת הטרנספורמר; הוא מעשיר את הבנתנו, ומעגן אותו במסורת מתמטית מאחדת החורגת מתחום ספציפי אחד. הוא מרמז שהעיקרון המרכזי, חישוב מבוסס-אפיניטי, רחוק מלהיות ממוצה**.**

https://arxiv.org/abs/2507.14560