המאמר היומי של מייק: 08.08.25  
Efficient Attention Mechanisms for Large Language Models: A Survey

מנגנון self-attention הוא הלב הפועם של מודלי שפה מודרניים. הוא מעניק לטרנספורמרים את יכולתם העמוקה להבין הקשר על ידי כך שהוא מאפשר לכל טוקן לתקשר עם כל טוקן אחר ברצף. אך לכוח הזה יש מחיר אדיר, כמעט בלתי אפשרי. דרישות החישוב והזיכרון של קשב עצמי גדלות באופן ריבועי ביחס לאורך רצף הקלט. צוואר בקבוק יחיד זה הגדיר במשך שנים את האופק של מה שאפשרי, והפך חשיבה בהקשר ארוך באמת לאתגר גדול.

נעשו מאמצים מחקרים משמעותיים במטרה להתמודד עם "הסיבוכיות הריבועית", שהולידו מספר רב של פתרונות מגוונים ולעיתים קרובות מבלבלים . סקירה של תחום זה לא רק מפרטת את השיטות הללו; היא מספקת טקסונומיה חיונית ומהווה מפה לניווט בין הפשרות המורכבות של יעילות חישובית, expressiveness של המודל ואלגנטיות תיאורטית. סקירה זו צוללת לעומק העקרונות המרכזיים המבנים את התחום הזה.

#### 4 משפחות היעילות

בבסיסו, האתגר הוא לקרב את מטריצת ה-attention המלאה בגודל N על N, מבלי לחשב או לאחסן אותה במפורש. הסקירה מסווגת את שלל הגישות לארבע משפחות עיקריות, שלכל אחת מהן פילוסופיה משלה.

1. דלילות בתבנית קבועה: התיקון הארכיטקטוני

הגישה הישירה ביותר לשבירת צוואר הבקבוק הריבועי היא להניח שמטריצת attention צפופה של "הכל-להכל" היא מוגזמת. שיטות אלו כופות תבנית attention דלילה וקבועה מראש, שבה כל טוקן רשאי להתייחס רק לתת-קבוצה קטנה וקבועה של טוקנים אחרים.

משפחה זו כוללת שיטות המשתמשות sliding windows, שבהן טוקן מתייחס רק לשכניו המקומיים. גישה זו מבוססת על האינטואיציה החזקה של "מקומיות ההקשר" (locality of reference) שמילים סמוכות הן לרוב הרלוונטיות ביותר. כדי למנוע אובדן של מידע גלובלי, גישה זו מחוזקת לעיתים קרובות באמצעות מספר טוקנים גלובליים הרשאים להתייחס לכל הרצף, או באמצעות תבניות מורחבות/מדלגות (dilated/strided patterns) המדלגות באופן שיטתי על טוקנים כדי לכסות שדה קליטה רחב יותר עם מספר קבוע של חישובים.

שיטות אלו יעילות מאוד ופשוטות ליישום, אך מגבלתן העיקרית היא הנוקשות שלהן. תבניות ה-attention מהונדסות ידנית ואינן תלויות בנתונים, מה שאומר שהמודל אינו יכול להחליט באופן דינמי להתמקד בטוקן מרוחק אך רלוונטי מחוץ לחלון שנקבע לו מראש.

2. קירוב מדרגה נמוכה (Low-Rank): טריק הדחיסה

משפחה זו של שיטות פועלת על בסיס תובנה מתמטית עדינה יותר: שמטריצת ה-attention המלאה היא לרוב מדרגה נמוכה, כלומר ניתן לדחוס ביעילות את המידע שבה למספר קטן בהרבה של "מושגים" או וקטורי סיכום. במקום לחשב את המטריצה המלאה, מודלים אלו מקרינים את מטריצות השאילתה, המפתח והערך (Query, Key, Value) לתת-מרחב בעל ממד נמוך יותר, ובכך מאלצים את מנגנון ה-attention לפעול דרך צוואר בקבוק של מידע.

הרעיון המרכזי הוא לקרב את מטריצת ה-N על N על ידי פירוקה למכפלה של שתי מטריצות קטנות יותר, בגודל N על k, כאשר k קטן משמעותית מ-N. במהות, המודל לומד לסכם את כל הרצף למספר קבוע של צמדי מפתח-ערך מייצגים, וכל הטוקנים מתייחסים לסיכום דחוס זה במקום זה לזה. זוהי גישה גמישה יותר מתבניות קבועות, שכן תוכן הסיכום הדחוס נלמד מהנתונים. עם זאת, הדבר מציג פשרה חדשה: הגודל הקבוע של צוואר הבקבוק מגביל את קיבולת המודל להתמודד עם רצפים בעלי צפיפות גבוהה מאוד של מידע ייחודי.

3. קרנליזציה (Kernelization): תעלול מתמטי קליל

אולי הפתרונות האלגנטיים ביותר מבחינה מתמטית הם אלו הממסגרים מחדש את ה-attention דרך עדשת שיטות הקרנל (kernel methods). ניתן לראות את ה-attention הסטנדרטי כתהליך של חישוב מטריצת דמיון בין שאילתות למפתחות, ולאחר מכן שימוש במטריצה זו כדי לשקלל את הערכים. העלות הריבועית נובעת מהבנייה המפורשת של מטריצת דמיון מסיבית זו.

שיטות מבוססות קרנל עוקפות זאת בתחכום על ידי מינוף התכונה האסוציאטיבית של כפל מטריצות. הן מנסחות מחדש את חישוב ה-attention כך שישלבו תחילה את המפתחות והערכים, לפני האינטראקציה עם השאילתות. שינוי סדר הפעולות הפשוט הזה מונע את יצירת המטריצה בגודל N על N. במקום מכפלת מטריצה-במטריצה גדולה, החישוב מצטמצם לשתי מכפלות קטנות יותר של מטריצה-בווקטור, מה שמוריד את הסיבוכיות מריבועית ללינארית.

גישה זו חזקה מכיוון שבתיאוריה, היא יכולה לקרב את מנגנון ה-attention המלא מבלי לכפות אילוצי דלילות נוקשים. יעילותה תלויה במציאת פונקציית קרנל שתופסת במדויק את הדמיון בין שאילתות למפתחות, וחלק גדול מהמחקר בתחום זה מתמקד בפיתוח פונקציות קרנל חדשות (לרוב באמצעות טכניקות כמו קירוב תכונות אקראי) שהן גם יעילות וגם בעלות יכולת ביטוי גבוהה.

4. דלילות נלמדת ו-Mixture of Experts: הגישה האדפטיבית

משפחה רביעית ומתפתחת שואפת להשיג את הטוב מכל העולמות על ידי הפיכת תבנית הדלילות עצמה לתלוית-נתונים ונלמדת. במקום להשתמש בתבניות קבועות או בצוואר בקבוק גלובלי מדרגה נמוכה, שיטות אלו מנסות לחזות אילו טוקנים הם הרלוונטיים ביותר עבור שאילתה נתונה. הדבר מושג לעיתים קרובות באמצעות טכניקות כמו אשכולות (clustering) או על ידי שימוש במסגרת MoE, שבה ראשי attention שונים מאומנים כ"מומחים" לסוגים שונים של תבניות. מנגנון ניתוב לומד לשלוח כל טוקן לראש המומחה הרלוונטי ביותר. גישות היברידיות אלו הן בין החזקות והגמישות ביותר, אך גם המורכבות ביותר ליישום ואימון.

לסיכום, הסקירה חושפת שאין מנגנון attention יעיל אחד שהוא "הטוב ביותר". כל משפחה מציגה בחירה מהותית לגבי אופי הקירוב, ומבצעת פשרה שונה בין סיבוכיות חישובית לכוח ביטוי. התחום הוא דיאלוג תוסס בין הנחות יסוד ארכיטקטוניות, תורת הקירוב המתמטית ומערכות אדפטיביות ולומדות.

https://arxiv.org/abs/2507.19595