**שומרי הסף של ה-attention: תיקון פשוט להפליא לבעיה יסודית ב-LLMs: סקירה 516**

**המאמר היומי של מייק: 26.09.25  
Gated Attention for Large Language Models- Non-linearity, Sparsity, and Attention-Sink-Free**

במלחמה המתמדת על עליונות בעולם ה-LLM, התקדמות בדרך כלל נראית כמו עוד: עוד נתונים, עוד פרמטרים, עוד כוח חישוב. התרגלנו להאמין שפריצות דרך הן תוספות מורכבות ובארוקיות לארכיטקטורה שכבר מתנשאת לגובה. זו הסיבה שהמאמר המסוקר מרגיש כמו משב רוח רענן. הוא מציג ממצא שהוא בו-זמנית פשוט עד כדי גיחוך ועמוק מבחינה רעיונית: הוספת מנגנון gating פשוט במקום הנכון בתוך בלוק הקשב לא רק משפרת את הביצועים באופן הדרגתי; היא משנה באופן משמעותי את זרימת המידע ומתקנת פתולוגיות שחשבנו שהן אנדמיות לארכיטקטורה.

זהו אינו מאמר על מודל חדש ונוצץ. זוהי עבודה של התבוננות ארכיטקטונית מעמיקה. הוא שואל שאלה פשוטה: מה קורה אם נוסיף מנגנון gating לפלט של קשב ה-softmax? והתשובה חושפת חולשות קריטיות בטרנספורמר הסטנדרטי.

#### **צוואר הבקבוק החבוי בכל ראש Attention**

כדי להעריך את תרומת המאמר, עלינו לחזור ולבחון מה ראש attention באמת עושה, לא כסדרה של כפלי מטריצות, אלא כסדרת טרנספורמציות. מנגנון ה-Attention מחשב לכל טוקן ייצוג חדש באמצעות שקלול לינארי של כל שאר טוקנים במשפט,(המיוצג באמצעות מטריצת value) כאשר משקל השקלול נקבע לפי מידת הרלוונטיות ביניהן.

כאן טמונה החולשה הנסתרת: הטרנספורמציה מהווקטורים המצטברים של ה-"Value" לפלט הסופי היא רצף של 3 פעולות לינאריות(הכפלות ב-V, שקלול ומכפלה במטריצה W\_o אחרי שרשור פלט ראשי ה-attention). מנקודת מבט מתמטית, שתי העתקות לינאריות רצופות יכולות להצטמצם לאחת, פחות אקספרסיבית. זה יוצר צוואר בקבוק "מדרגה נמוכה" (low-rank). ראש attention בעצם נאלץ לדחוס את כל המידע העשיר והמשוקלל מבחינה הקשרית שהוא זה עתה חישב, דרך צוואר בקבוק של מידע.

#### **הפתרון: אי-לינאריות ודלילות (Sparsity)**

ההתערבות שהמחברים מציעים פשוטה להפליא: החלת מנגנון gating לא לינארי אחרי השרשור של כל תוצאות ראשי ה-attention. פעולת הכפל הפשוטה הזו עושה שני דברים חיוניים.

ראשית, היא מזריקה מנה שלאי-לינאריות בדיוק במקום שהיא הכי נחוצה. על ידי שבירת רצף הפעולות הלינאריות, השער מנפץ את צוואר הבקבוק מדרגה נמוכה. צעד יחיד זה פותח רמה גבוהה יותר של אקספרסיביות עבור ראש ה-attentiot, ומאפשר לו למדל קשרים מורכבים הרבה יותר בין טוקנים. הוא כבר לא רק מסובב ומותח מידע; כעת הוא מסוגל לקבל החלטות חדות ולא-לינאריות לגביו.

שנית, וזו התובנה העמוקה יותר, השער מציג דלילות תלוית-שאילתה (query-dependent sparsity). ערכי השער אינם קבועים; הם מחושבים על בסיס האמבדינג של הטוקן הנוכחי (השאילתה). משמעות הדבר היא שעבור כל טוקן המעובד, כל ראש attention "לומד להחליט" באופן דינמי אילו חלקים מהפלט של עצמו אינם רלוונטיים ויש "להנמיך" או לאפס אותם לחלוטין. זהו מסנן חכם ומודע לתוכן שגוזם מידע חסר תועלת *לאחר* שהוא כבר נאסף. זה יוצר אות דליל ונקי המועבר לשכבה הבאה של הרשת.

#### **התרופה לתופעה מסתורית: "בולען הקשב" (Attention Sink)**

ההשפעה השנייה הזו, דלילות דינמית, מתגלה כתרופה לפתולוגיה מתועדת היטב אך לא מובנת היטב של LLMs: "בולען האטנשן" (attention sink). מודלים חזקים רבים מציגים נטייה מוזרה להקצות כמות גבוהה באופן לא פרופורציונלי של attention לטוקן הראשון בסדרה (לרוב טוקן BOS או תחילת הסדרה), ללא קשר לרלוונטיות שלו. טוקן זה פועל כמו מזבלה חישובית, מקום שפונקציית ה-softmax יכולה לשלוח אליו ציוני attention שאין להם לאן ללכת. זהו תוצר לוואי בזבזני ולא יעיל של הארכיטקטורה.

המחברים מגלים שהגישה שלהם "חופשיה מבולען קשב" ("attention-sink-free"). מדוע? מכיוון שהשער מספק מנגנון ישיר ויעיל הרבה יותר להתעלמות מהקשר לא רלוונטי. במקום ללמוד לזרוק ציוני attention לא רצויים על טוקן בולען במהלך חישוב ה-softmax, המודל יכול כעת להמתין עד *לאחר* שהמידע נאסף ופשוט להשתמש בשער כדי לאפס את מה שהוא לא צריך. השער תלוי-השאילתה הופך את ה-attention sink למיושן.

מעבר לתיאוריה, לשינוי זה יש השפעות מעשיות חזקות. על ידי הפיכת פלט הattention לדליל וריסון הערכים המספריים הפראיים הידועים כ"האקטיבציות עצומות" (massive activations), השער מייצב באופן משמעותי את האימון. זה מאפשר שיעורי למידה גבוהים יותר ותכונות סקיילינג טובות יותר, מה שהופך את המודלים לא רק ליותר ביצועיסטיים אלא גם ליותר חסונים לאימון.

מאמר זה מזכיר לנו שהדרך קדימה לא תמיד כרוכה בבניית משהו חדש, אלא בהבנה עמוקה, ותיקון, של היסודות שכבר יש לנו. הוא מוכיח שלפעמים, המהלך החזק ביותר הוא לא להוסיף מורכבות, אלא להציג אילוץ יחיד ואלגנטי המאפשר למערכת כולה להתארגן בצורה חכמה יותר.

https://arxiv.org/abs/2505.06708