**הגיאומטריה של המיקוד: פירוק מגבלות ה-attention, סקירה 525**

**סקירת המאמר היומית של מייק: 13.10.25  
LIMITATIONS OF NORMALIZATION IN ATTENTION MECHANISM**

קיבלתי לא מזמן בקשה לסקור מאמרים יותר תאורטיים והנה מגיע לכם אחד כזה (בערך).

קהילת הלמידה העמוקה התייחסה זמן רב למגבלות של מנגנון attention סטנדרטי (עם softmax) כאל כורח המציאות שהחוקרים מנסים לתקן באמצעות פלסטרים ארכיטקטוניים (כגון נרמולים בבלוק הטרנספורמר). אנו רואים שככל שאורך ההקשר גדל, ה-attention "מיטשטש" ומאבד את יכולתו להתמקד. אנו יודעים שחידוד אגרסיבי של ההתפלגות באמצעות טמפרטורות נמוכות עלול לערער את יציבות האימון. אך התשובה לשאלת ה"למה" הבסיסית ניתנה לרוב באמצעות אינטואיציה ולא הוכחה פורמלית. מאמר זה משנה זאת, ומחליף היוריסטיקות בגבולות מתמטיים מוצקים. החידוש שלו אינו מנגנון קשב חדש, אלא מסגרת תיאורטית רבת עוצמה להבנת המגבלות המהותיות והבלתי נמנעות של *כל* סכמת נורמליזציית קשב.

החידוש המרכזי הראשון הוא הפורמליזציה של המושג "קיבולת בחירה". דמיינו שתפקידו של הקשב הוא לבחור N טוקנים חשובים מתוך הקשר באורך L. המאמר גוזר חסם עליון אסימפטוטי על המרחק הייצוגי בין הטוקנים ה"נבחרים" המאוגדים לבין אלו ש"לא נבחרו". זו אינה רק תצפית; זוהי הוכחה. המסגרת מדגימה שככל שמספר הטוקנים הנבחרים (N) גדל באופן יחסי לאורך ההקשר (L), יכולתו של המודל להבחין בין שתי הקבוצות מובטחת מתמטית לקרוס. המרחק ביניהן חייב להצטמצם, ובכך למעשה לטשטש את הגבול בין אות לרעש. הדבר מוכיח באלגנטיות את קיומו של "צוואר בקבוק סופטמקסי" מעקרונות ראשוניים, ומראה שאין זה פגם באופטימיזציה אלא תכונה בסיסית של הנורמליזציה עצמה.

התרומה השנייה של המאמר, ואולי האלגנטית ביותר, היא פרשנות גיאומטרית חדשנית למגבלות ה-attention. הוא ממפה מחדש את שיכוני הטוקנים כנקודות הפרוסות על פני ספירה(sphere) רב-ממדית ושואל שאלה פשוטה: כמה מתוך N הטוקנים המובילים שנבחרו יכול ראש attention יחיד באמת לשמור במיקוד בבת אחת? המחברים גוזרים חסם גיאומטרי המכמת תקרה קשיחה ל"יכולת הפרדה" זו. המאמר מוכיח שגם תחת הנחות מרחביות אידיאליות, ראש יחיד אינו יכול להבחין בו-זמנית בכל הטוקנים שהוא בוחר. חלק מהם יימצאו בהכרח מחוץ לאזור הניתן להבחנה, ויתמזגו עם הרקע. הדבר מספק טיעון גיאומטרי מדויק מדוע attention מרובה-ראשים אינו רק מועיל אלא הכרחי: לא רק כדי לתפוס דפוסים לשוניים מגוונים, אלא כדי להתגבר על הרזולוציה הסופית של נקודת מבט גיאומטרית יחידה.

לבסוף, העבודה מכלילה את הפשרה הידועה בין חדות ה-attention ליציבות הגרדיאנט. בעוד שאנו יודעים שהורדת טמפרטורת הסופטמקס עלולה להוביל לגרדיאנטים תנודתיים, המחברים מרחיבים זאת לעיקרון אוניברסלי. הם גוזרים חסם יעקוביאני עבור סוג יחסית כללי של פונקציות נרמול (כמו סופטמקס), ומוכיחים ש*כל* פונקציה המייצרת התפלגות חדה וסלקטיבית תסבול באופן מובנה מרגישות לגרדיאנט גבוהה. הדבר ״מרומם״ טריק ידוע הקשור לסופטמקס למעמד של חוק יסוד במכניקת ה-attention, ומסביר את הפשרה הבלתי נמנעת בין סלקטיביות ליכולת אימון.

https://arxiv.org/abs/2508.17821