המאמר היומי של מייק - 09.01.25  
When Can Transformers Count to n?

המאמר חוקר את המגבלות התיאורטיות והאמפיריות של ארכיטקטורות טרנספורמר כאשר בביצוע משימות ספירה פשוטות. הוא בוחן משימות כמו "ספירת שאילתות" (QC) ו"האלמנט השכיח ביותר" (MFE) כדי לקבוע מתי טרנספורמרים יכולים לפתור בעיות אלה ביעילות. המחקר חושף הן את היכולות והן את המגבלות המובנות של טרנספורמרים בהקשרים כאלה, ומספק תובנות מעניינות לגבי האילוצים הארכיטקטוניים שלהם.

התרומות העיקריות:

משימת QC

משימת QC היא למעשה ספירה של כמה פעמים טוקן מסוים מופיע ברצף. המחברים מדגימים שהטרנספורמרים יכולים לבצע משימה זו ביעילות אם גודל האמבדינג d גדול מפי שניים מגודל המילון m: עבור d > 2m, גישה מבוססת היסטוגרמה (שפותחה במאמר) מאפשרת ספירה על ידי הטמעת ייצוגי טוקנים בצורה אורתוגונלי. זה מאפשר למודל לבנות היסטוגרמה של הופעות טוקן על ידי בלוק טרנספורמר יחיד. עבור d < m, האורתוגונליות של המבדינגס כבר לא אפשרית, מה שהופך ספירה מדויקת לבלתי אפשרית. המאמר מוכיח מגבלה זו בקפידה באמצעות חסמי Welch, המאפשר לנתח את את הטרייד-אופים של מימד האמבדינגס (הקשורים לאורתוגונליות).

שיטת CountAttend

כאשר גודל אמבדינגס d קטן מגודל המילון m, המחברים מציעים את פתרון ה-"CountAttend", כדי לפתות את QC עם מנגנוני ה-attention. הפתרון כולל שני רכיבים עיקריים:

1. משקלי attention

- מנגנון attention מייצר משקלים המקודדים את היחס בין טוקן השאילתה לכל הטוקנים ברצף. לצורך ספירה, משקלי attention חייבים להיות הפוכים ביחס לספירת הטוקן ברצף.שקלול זה מבטיח שתרומת כל אסימון לפלט מנורמל לפי התדירות של - זה מבטיח שתרומה של כל טוקן לפלט מנורמלת לפי התדירות שלו.

2. MLP להיפוך משקלים

- נדרש MLP כדי לשחזר את הספירה האמיתית c ממשקלי attention. ה-MLP צריך ללמוד פונקציה מהצורה: f(w) = 1/w כאשר w הוא משקל attention עבור טוקן

אתגרים עם פתרון CountAttend:

חישוב משקלי attention: חישוב משקלים הפוכים ביחס לספירות טוקנים דורש מידול מדויק של יחסי טוקנים לאורך הרצף. זה מוסיף מורכבות למנגנון ה-attention

גודל MLP: עבור סדרות ארוכות יותר, מספר הנוירונים ב-MLP חייב לגדול באופן פרופורציונלי ביחס לאורך הסדרה שזה בעייתי מאוד מבחינה חישובית.

משימת MFE:

מטרת משימת ה-MFE, היא למצוא טוקן בעל התדירות הגבוהה ביותר בסדרה. ניתן ליישם את המשימה אם (d=O(m באמצעות גישה מבוססת היסטוגרמה. עבור d < m, המשימה הופכת לבלתי אפשרית, כפי שמוכח באמצעות טיעוני מורכבות תקשורת. המחברים מציעים פתרון טרנספורמר דו-שכבתי לבעיה זו.

מעבר פאזה בביצועים

המאמר מזהה מעבר פאזה קריטי: טרנספורמרים נכשלים במשימות ספירה כאשר d < m. סף זה מדגיש את הפשרות בין גודל ההטמעה, גודל אוצר המילים, ומורכבות המשימה.

תובנות תיאורטיות:

בניית אמבדינגס אורתוגונליים

המחברים מנצלים את התכונות המתמטיות של א אורתונורמליות ליישום ספירה מבוססת היסטוגרמה. עבור d > m, ניתן לבנות אמבדינגס כך שמכפלה סקלרית בין הטמעות טוקן שונות היא אפס. זה מבטיח ספירת טוקנים מדויקת בתוך בלוק attention יחידה. המאמר משתמש בגבולות Welch להראות שעבור d < m, המכפלה הפנימית בין וקטורי ההטמעה הופכת משמעותית, מה שמכניס שגיאות בהיסטוגרמה. עבור משימת MFE, המחברים משתמשים בכלים מעולם communication complexity כדי להוכיח שהטרנספורמרים דורשים (d=Ω(m כדי לפתור את המשימה.

השלכות מעשיות:

המסקנות מובילות למספר השלכות לתכנון ופריסה של טרנספורמרים ביישומים מעשיים: סקלביליות ארכיטקטונית: טרנספורמרים חייבים להתאים את גודל אמבדינג לגודל המילון. קידוד מיקומי (positional encoding) המחברים מדגישים את הנחיצות של הטמעות מיקום למשימות ספירה. בעוד שפתרון ההיסטוגרמה יעיל עבור d > m, היישום המעשי שלו עשוי להיות מאוד בעייתי מבחינת הזכרון וסיבוכיות.

מסקנה

המאמר מספק ניתוח מקיף של היכולות והמגבלות של טרנספורמרים בפתרון משימות ספירה בסיסיות. באמצעות שילוב של הוכחות תיאורטיות ריזורוזיות עם אימות אמפירי, הוא מדגיש את הפשרות הארכיטקטוניות המובנות במודלי הטרנספורמרים.

מחקר עתיד:

- ארכיטקטורות היברידיות המשלבות טרנספורמרים עם שיטות ניורו-סימבוליות למשימות ספירה

- הרחבות למשימות הכוללות ספירה היררכית או מובנית

- מחקרי mechanistic interpretability להבהרת הייצוגים הפנימיים שנלמדים על ידי טרנספורמרים במהלך משימות ספירה

https://arxiv.org/pdf/2407.15160