🚀המאמר היומי של מייק -03.10.24: ⚡️🚀  
Transformers are Expressive, But Are They Expressive Enough for Regression?

שוב מאמר על הטרנספורמרים אבל קצת שונה מהמאמר הסטנדרטי על LLMs. המאמר הזה מציג חקירה מעמיקה לגבי expressiveness של הטרנספורמרים, תוך בחינה ספציפית של יכולתם בתור משערכי פונקציות אוניברסליים (כאלו שניתן לקרב איתם כל פונקציה חלקה בדיוק נתון). המחברים מאתגרים טענות קיימות לגבי expressiveness של הטרנספורמרים ומספקים הוכחות תיאורטיות ואמפיריות כאחד שתומכים בהשערתם שהטרנספורמרים מתקשים לקרב (לשערך) באופן מדויק פונקציות חלקות.

לפני 4 שנים הוכח שהטרנספורמר(האנקודר) מסוגל לשערך כל פונקציה רציפה אם יש בו מספיק שכבות (בלוקים של טרנספורמר). המשפט הוכח לפני כ 4 שנים והוא מראה שהטרנספורמר בעל שכבות מרובות למעשה יודע לשערך ופונקציה קבועה למקוטעין (piecewise constant) ועם הגודל המינימלי של אינטרוול הקביעות (=רזולוציה) δ הינו קטן מדי אז ניתן לשערך באמצעותו כל פונקציה חלקה בכל דיוק.

המאמר המסוקר מתמקד במחקר של הרזולוציה δ הנדרשת לשערוך בדיוק נתון של פונקציה חלקה. התרומה התיאורטית המרכזית של המאמר היא משפט 4.1, אשר קובע חסם עליון על גורם הרזולוציה δ עבור שמכיל מאפיינים שונים של פונקציה מקורבת f.

משפט זה משמעותי מכמה סיבות:

א) הוא קושר ישירות את גורם הרזולוציה δ לנגזרות של f. קשר זה מבהיר מדוע פונקציות חלקות עם נגזרות המשתנות במהירות מהוות אתגר קשה עבור טרנספורמרים.

ב) החסם מראה יחס הפוך בין δ לבין הנגזרות החלקיות של הפונקציה. עבור פונקציות עם נגזרות גדולות, δ חייב להיות קטן כדי לשמור על איכות הקירוב. זה אומר בעצם שאנו צריכים יותר שכבות של טרנספורמרים כדי לקרב בדיוק גבוה את f.

ג) המונח האקספוננציאלי 1/(p+md) בחסם מצביע על כך שככל שממד הקלט m או ממד האמבדינג d גדלים, גורם הרזולוציה δ חייב לקטון אקספוננציאלית כדי לשמור על אותה איכות קירוב.

ד״א המחברים מספקים הוכחה מפורטת למשפט זה, תחילה למקרה החד-ממדי ולאחר מכן בהכללה לממדים גבוהים יותר..

יתר על כן, המחברים מקשרים את התוצאה התיאורטית הזו להשלכות המעשיות על ארכיטקטורות טרנספורמר. הם מראים שמספר השכבות הנדרש לקירוב הולם גדל כ ((O(m(1/δ)^(dm, מה שהופך ללא ישים מבחינה חישובית עבור δ קטן וממד הקלט בגודל בינוני m. כלומר צריך יותר מדי שכבות הטרנספומרים בשביל זה.

המחברים ביצעו ניסויים מקיפים על הטרנספורמר כדי להשלים את ממצאיהם התיאורטיים. הם עשו 2 ניסויים עם הבנצ'מרקים הבאים:

א) EXPT-I (רגרסיה): בדיקת יכולתם של טרנספורמרים לקרב ישירות פונקציות חלקות.

ב) EXPT-II (״סיווג מקוונטט״): בדיקת יכולתם של טרנספורמרים לקרב פונקציות קבועות למקוטעין.

התברר כי הטרנספורמרים מתפקדים באופן גרוע משמעותית ב-EXPT-I בהשוואה ל-EXPT-II, שזה תומך בהשערה שהם מתקשים בקירוב פונקציות חלקות.

הגדלת מספר השכבות, ראשי מנגנון ה-attention, או ממדי  אמבדינג אינה משפרת באופן משמעותי את הביצועים על פונקציות חלקות. לעומת הטרנספורמרים מצליחים לקרב באופן הולם פונקציות קבועות למקוטעין עם רזולוציה δ לא קטנה במיוחד.

https://arxiv.org/pdf/2402.15478