המאמר היומי של מייק - 22.12.24:  
Reasoning in Large Language Models: A Geometric Perspective

מאמר זה חוקר את יכולות החשיבה של LLMs מנקודת מבט גיאומטרית, תוך התמקדות בקשר בין הממד הפנימי(intrinsic dimension או ID) של ייצוגי הקלט לבין עוצמת expressiveness של מודלים אלה. החוקרים בוחנים כיצד ארכיטקטורות טרנספורמר מחלקות את מרחבי הקלט וכיצד חלוקה זו קשורה ליכולות ההנמקה שלהן (reasoning). העבודה מציעה תובנות חשובות לגבי האופן שבו ארכיטקטורת המודל ואורך ההקשר משפיעים על ביצועי LLM במשימות הנמקה.

רעיונות מרכזיים:

מסגרת גיאומטרית לכוח ביטוי של מדוך (אקספרסיביות)  
הרעיון המרכזי סובב סביב צפיפות גרפי מנגנון self-attention והשפעתה על הממד הפנימי של הקלטים לשכבות MLP בתוך הטרנספורמרים (כלומר FFN). המימד הפנימי, בהקשר זה, מודד את מספר דרגות החופש האפקטיביות הנדרשות לייצוג אמבדינג של הקלט.

מנגנון self-attention כגרף:  
הפלט של שכבת מנגנון self-attention מתואר כגרף, בו טוקנים הם צמתים ומקדמי attention מגדירים קשתות משוקללות. צפיפות הגרף קובעת את מספר החיבורים האפקטיביים, המשפיעים ישירות על הממד הפנימי של הייצוגים המועברים לבלוקי MLP.

חלוקת מרחב הקלט:  
ממדים פנימיים גבוהים יותר מאפשרים לשכבות ה-MLP לחלק את מרחב הקלט לאזורים עדינים יותר. זה מאפשר למודל לבנות מיפויים מורכבים יותר ולתפוס קשרים לא-לינאריים ביעילות. כתוצאה מכך, יכולת ההנמקה של ה-LLM משתפרת עם כוח הביטוי המוגבר הנובע מחלוקות עדינות אלה.

יכולות קירוב:  
על ידי אפשור חלוקה עדינה יותר, ממדים פנימיים גבוהים יותר מפחיתים שגיאות קירוב, מאפשרים ל-MLP לייצג פונקציות מורכבות בדיוק רב יותר. זה מתקשר ישירות למשימות הנמקה, בהן מיפויים מדויקים ותלויי הקשר הם קריטיים.

הסברים מעמיקים על הרעיונות:

חלוקה וקירוב

החוקרים משתמשים בניסוח piece-wise affine של רשתות נוירונים עמוקות (DNNs) כדי להסביר כיצד מרחב הקלט מחולק. הרעיון המרכזי של חלק "החלוקה והקירוב" הוא לתאר כיצד DNNs מחלקות את מרחב הקלט למספר אזורים, כל אחד נשלט על ידי כלל ליניארי ספציפי משלו.

חלוקת מרחב הקלט:

רשתות נוירונים (באמצעות פונקציות אקטיבציה בשכבותיה), מחלקות את מרחב הקלט למספר אזורים מובחנים. אזורים אלה מוגדרים על בסיס האופן שבו הנוירונים מופעלים בתגובה לנתוני הקלט. חשבו על מרחב הקלט כמפה, והרשת יוצרת "אזורים" על מפה זו כאשר לכל אזור יש כלל ייחודי משלו.

קירוב לינארי בתוך כל אזור:

בתוך כל אזור כזה, הרשת מתנהגת כמו פונקציה לינארית. זה למעשה מאפשר קירוב פונקציות מורכבות יותר על ידי שילוב חלקים פשוטים אלה.

יכולת הרשת לקרב פונקציות מורכבות תלויה ביכולתה לחלק את מרחב הקלט ו"להגדיר" חוקים לכל אזור. יותר חלוקות מאפשרות קירוב טוב יותר, שהוא קריטי למשימות מורכבות כמו הנמקה. מסגרת זו עוזרת להבין כיצד רשתות משתמשות באבני בניין פשוטות (מודלים לינאריים באזורים ספציפיים) כדי להתמודד עם בעיות מורכבות מאוד. ניסוח זה מדגיש את היכולת של DNNs לחלק באופן אדפטיבי את מרחב הקלט על בסיס דאטה האימון, כאשר מספר האזורים מתואם ישירות עם כוח הקירוב של המודל.

עבור טרנספורמרים, תורה זה ניתנת ליישום למנגנון multi head self attention או MHST שבו צפיפות האינטראקציות בין טוקנים משפיעה על החלוקה המושרית של מרחב הקלט ברמת שכבות ה-MLP שלו.

משפט מרכזי:   
סכום מינקובסקי מסביר כיצד הפלטים של שכבת MHST מובנים גיאומטרית וקשורים למושג המימד הפנימי. בטרנספורמרים, MHST מפצלת את מנגנון attention למספר "ראשים", כאשר כל ראש מתמקד בהיבט ספציפי של הקלט. ראשים אלה עובדים במקביל כדי לתפוס יחסים שונים בתוך הדאטה. המשפט מראה ניתן לפרש את הפלט של MHST כשילוב של אזורים שנוצרו על ידי כל ראש בודד. כל ראש מגדיר "צורה" (טכנית, מעטפת קמורה) המבוססת על הטרנספורמציות שהוא מחיל על הקלט.

סכום מינקובסקי:  
סכום מינקובסקי הוא פעולה מתמטית המשמשת לשילוב צורות אלה. באופן אינטואיטיבי, זה אומר שהפלט של שכבת MHST הוא מרחב הכולל את כל השילובים האפשריים של פלטי הראשים הבודדים.

קשר למימד פנימי:  
תוצאה זו מדגישה שהוספת ראשים נוספים או הפיכת הראשים לאקספרסיביים יותר מגדילה את ה"ממדיות" של המרחב שבו נמצאים פלטי תשומת-הלב. ממדיות מורחבת זו משפרת את יכולת המודל לייצג יחסים מורכבים ותהליכי חשיבה. המשפט מפרמל כיצד מנגנון MHST מחלק ומשלב את ההיבטים הגיאומטריים של מרחב קלט כדי להגביר את האקספרסיביות ויכולת הנמקה של מודלי טרנספורמר.

המימד האפקטיבי של סכום מינקובסקי תלוי בצפיפות גרף attention (כלומר, מספר החיבורים הפעילים בין טוקנים). צפיפות גרף גבוהה יותר, המושגת באמצעות יותר ראשי attention או קישוריות גבוהה יותר, מובילה לממדיות פנימית גדולה יותר של הקלט לשכבות MLP. מימד פנימי בוחן עד כמה טוב טרנספורמר יכול לתפוס יחסים מורכבים בקלט שלו בהתבסס על מספר החיבורים המשמעותיים שהוא מזהה.

מימד פנימי גבוה יותר פירושו שיותר חלקים מהקלט משפיעים על טוקן. זה מוביל לייצוגים עשירים ומפורטים יותר של הקלט, המאפשרים למודל להבין טוב יותר דפוסים ויחסים מורכבים. כאשר למודל יש ממד פנימי גבוה, הוא יכול "לחלק" ביעילות את מרחב הקלט ליותר אזורים, מה שמאפשר לו לתפוס פרטים ודקויות עדינים יותר. זה קריטי למשימות חשיבה, שבהן הבנת יחסים עדינים היא מפתח.

השלכות מעשיות:  
הגדלת מספר ראשי attention או קלטים ארוכים יותר עשויים להגדיל את המימד הפנימי. זה משפר את יכולות החשיבה של המודל מבלי לדרוש שינויים בארכיטקטורה שלו או בתהליך האימון. הממד הפנימי משקף עד כמה עמוק טרנספורמר מתעסק עם הקלט שלו. ככל שהחיבורים עשירים יותר, כך המודל יכול לחשוב טוב יותר ולבצע משימות מורכבות.

https://arxiv.org/abs/2407.02678