⚡️🚀המאמר היומי של מייק -22.11.24: ⚡️🚀

The Unreasonable Ineffectiveness of the Deeper Layers

מאמר קליל שלא יקשה עליכם יותר מדי בסופ״ש. המאמר מציע דרך מאוד פשוטה לקצץ שכבות במודלים המבוססים על ארכיטקטורת הטרנספורמרים. אתם בטח זוכרים שמודלי שפה שלנו וגם לא מעט מודלים בדומיינים אחרים מבוססים על טרנספורמרים שמורכבים מבלוקים שכל אחד מהם מורכב ממנגנון attention ושתי שכבות feed-forward (השנייה מהן לינארית). בנוסף יש שכבות נרמול וחיבורי residual (כלומר הפלט של כל שכבה מחובר יחד עם הפלט של השכבה הקודמת).

מודלי שפה מודרניים מכילים עשרות רבות של בלוקי טרנספורמרים שכמובן משליך על כמות הזמן והמשאבים הנדרשים להפעלתם, בעיקר במשימות גנרוט. כאמור המאמר שנסקור היום מציע דרך לקצץ כמה בלוקי טרנספורמרים רצופים שכמובן יקטין את זמן חישוב שנדרש ליצירה הפלט. אבל איזה בלוקים לבחור כך שהפגיעה בדיוק המודל תהיה מינימלית.

מכיוון שהגרף החישובי של הטרנספורמר מורכב מלא מעט חיבורי residual טבעי לבחור בלוקים רצופים שלא מוסיפים הרבה לפלט הבלוק הנמצא לפניהם במודל. כלומר אם הדלתא שנותנים הבלוקים האלו זניחה אז ניתן להעיף אותם בלי פגיעה רצינית בביצועים.

האבל איך ניתן לבדוק את זה? האמת יש לא מעט דרכים לבדוק את זה ומאמר בחר להשוות את הפלט של הבלוק l עם הפלט של הבלוק l+n (אנו מוחקים n בלוקים רצופים) באמצעות מודיפיקציה קטנה של מרחק קוסיין (החליפו cos ב-arccos וחילקו ב-pi כדי לגרום למדד הזה להיות בין 0 ל 1). באופן הגיוני n בלוקים עם דמיון גבוה מאוד לבלוק שקודם להם (מבחינת הפלט) נבחרים בתור מועמדים טובים לקיצוץ (כלומר בוחרים בלוק התחלתי l ומספר בלוקים לקיצוץ n עם הדמיון הגבוה ביותר). הדמיון מחושב על ייצוג הטוקן האחרון עבור כמות דאטה גדולה.

לאחר המחיקה ניתן לעשות למודל פיין טיון קליל ולטענת המחברים ניתן למחוק ככה על חצי שכבות טנרספורמים (במודלי שפה) בלי פגיעה רצינית בביצועים).

https://arxiv.org/abs/2403.17887