⚡️🚀המאמר היומי של מייק -10.11.24: ⚡️🚀

WHAT MATTERS IN TRANSFORMERS? NOT ALL ATTENTION IS NEEDED

סקירה קצרה של מאמר די נחמד החוקר איזה חלקים במודלי טרנספורמרים (או שנאים) שלנו פחות נחוצים מהחלקים האחרים (או בכלל מיותרים). כמו שאתם זוכרים בכל בלוק של שנאי יש לנו מנגנון ה-attention, כמה שכבות MLP (שזה שכבה וחצי של fully-connected) וכמה שכבות נרמול (אותם לא בודקים).

המחקרים הקודם שחקרו את הנושא הזה התמקדו בזיהוי בלוקים שלמים של שנאים העשויים להיות לא נחוצים אך המחקר הזה החליטו לרדת לרזולוציה של אבן הבניין של השנאי עצמו (כלומר attention ו-MLP).

איך בודקים האם תת-בלוק לא נחוץ? בודקים את הקלט את הפלט של תת הבלוק הזה ואם אין כמעט הבדל בינם כנראה שלא צריך אותו. כדי לבדוק את הדמיון משתמשים כמובן בדמיון קוסיין (cosine similarity). בודקים את זה על כמויות גדולות של דאטה ומתחילים להוריד שכבות ולבדוק ביצועים.

מה התברר? באופן קצת מפתיע לרוב מנגנוני ה-attention הרבה פחות נחוצים מה-MLP וניתן לוותר עליהם בלי פגיעה רצינית בביצועים במיוחד במודלים הגדולים. אז אולי זו הדרך להקטין את העומס החישובי ששימוש המודלים האלו גורם? בואו נחכה ונראה….

https://arxiv.org/abs/2406.15786