המאמר היומי של מייק - 24.03.25  
STAR ATTENTION: EFFICIENT LLM INFERENCE OVER LONG SEQUENCES

הסקירה הזו הולכת להיות קצרה. אפילו מאוד קצרה. המאמר המסוקר מציע שיטה לאופטימיזציה של מנגנון ה-attention בטרנספורמרים עבור מקרה שיש לנו כמה מכונות (נקרא hosts במאמר) להריץ את מודל השפה שלנו. המאמר הוא של חברת אנוידיה דרך אגב וזה דווקא לא מפתיע כי (לפי השמועות 🙂) יש להם די הרבה משאבי חישוב.

המודל מחזיר אותי לתקופה העליזה מלפני 4-7 שנים שהייתי עד למבול של מאמרים שהציעו אופטימיזציות שונות למנגנון ה-attention. אתם בטח זוכרים LongFormer, Performer, Reformer, LinFormer וכדומה(שחלקם סקרתי בזמנו) - היה גם Star Transformer דרך אגב. רוב השכלולים שהוצעו בתקופה ההיא דיברו על איך ניתן לזרז את ה-attention בלי לפגוע משמעותית בביצועי המודל - כאשר המודל רץ על מכונה אחת. אז היה מאוד פופולרי האיורים הריבועיים שהיה מצויר בהם הפאטרן של ה-attention כלומר באיזה טוקנים טוקן נתון מתחשב כדי לבנות את ייצוגו ההקשרי (contextualized embedding).

המאמר הזה מציע מנגנון attention שניתן לקרוא לו לוקאלי (מזכיר לי קצת רשתות קונבולוציה על ה-inductive bias שלהם המנצל את התלויות הלוקאלית בתמונות). במאמר זה משהו טיפה יותר מורכב (מזכיר גם LongFormer). כאן מחלקים את חלון הקשר לכמה קבוצות של טוקנים c1,...c\_n. כל טוקן בכל קבוצה c\_i פרט ל-c1 מחשבת את ה-attention עם הטוקנים בתוך אותה הקבוצה ו-c1 בלבד כאשר טוקנים של c1 מתחשבים בכל הטוקנים לבניית האמבדינג שלהם. כלומר הקבוצה הראשונה של הטוקנים משפיעה על האמבדינגס של כל הטוקנים וגם בעצמה מושפעת מכל הטוקנים בחלון ההקשר. המחברים טוענים שללא הוספה של c1(שזה למעשה התחלת הפרומפט) לכל קבוצות הטוקנים המנגנון סובל מירידה רצינית בביצועים

כמובן ניתן למקבל את התהליך הזה בקלות בין כמה מכונות (hosts) כאשר כל host מחשב את ה-attention הלוקאלי שלו וגם ה-attention עם c1(בשני שלבים). כל host גם שומר את סכום האקספוננטים של Q ו-K (מכנה של הסופטמקס) עבור הטוקנים שלו. לאחר מכן כל הסכומים האלו מועברים ל-host נוסף שמנרמל את כולם עם סכום אקספוננטים של כל ה-hosts ומחשב את הייצוג הסופי של כל וטוקנים.

מנגנון זה מאפשר חישוב ממקובל ומהיר יותר של ה-attention (פחות מכפלות מטריצות) כאשר לטענת המחברים הפגיעה בביצועים לא משמעותית.

https://arxiv.org/abs/2411.17116