המאמר היומי של מייק - 06.03.25  
LYNX: ENABLING EFFICIENT MOE INFERENCE THROUGH DYNAMIC BATCH-AWARE EXPERT SELECTION

שמתי לב שמזמן לא סקרתי מאמר על MoE - Mixture Of Experts במודלי שפה. אזכיר ש-MoE זו שיטת המיועדת לאופטימיזציה של אינפרנס מבחינת העומס החישובי(כלומר פחות חישובים). המודל מאומן להפעיל רק חלק מהמודל (מומחים מסוימים) עבור כל טוקן כאשר כל מומחה הוא (בדרך כלל) תת-רשת של ה-FFN (למעשה תת-מטריצות של מטריצות המשקולות ב-FFN) בתוך מנגנון הטרנספורמרים. בפועל זה מאפשר להקטין את כמות החישובים לכל טוקן שעשוי לאפשר הפעלה של LLMs בגודל עצום (רק החלק מהמודל כל פעם). בנוסף (לפי כמה מחקרים) שיטה זו מאפשרת ללמוד ״פונקציות מורכבות יותר״ כי כל טוקן עשוי להיות מחושב בצורה שונה (עם תת-קבוצה שונה של מומחים).

המומחים נבחרים על ידי רשת ניתוב (routing network) כאשר היא מאומנת לחשב ציון אי שלילי לכל מומחה. ציונים הם למעשה ״הסתברות״ לבחירה של כל מומחה (יש softmax בסוף). בד״כ k מומחים בעלי ציונים הגבוהים ביותר נבחרים בכל שכבה עבור כל טוקן מתוך N מומחים כאשר k < N. המודל מאומן לאזן ניצול של כל מומחה כאשר המטרה שכל מומחה ינוצל במידה שווה בדאטהסט אימון (aggregative level). בד״כ יש איבר רגולריזציה על משקלי רשת הניתוב למשל בצורה של אנטרופיה שלילית או סכום הריבועים).

המאמר מציע שיטה לאופטימיזציה של צריכת זכרון עבור אינפרנס של מודלי טרנספורמרים עם MoE כאשר הם מופעלים בבאצים של שאילתות ( כמה קלטים). הגישה המוצעת מבוססת על כמה אובזרבציות אמפיריות שנעשו על ידי המחברים:

התפלגות של שכיחות הפעלת המומחים בתוך הבאץ' אינה אחידה כלומר יש מומחים שמופעלים יותר ויש כאלו שמופעלים פחות

הצפיפות החישובית (arithmetic intensity) שהיא היחס בין כמות הflops לכמות גישות זיכרון יורדת כאשר כמות המומחים עולה בשלב decode (כלומר חיזוי). זה הופך את השלב הזה ל-memory-bound שמגדיל את ה-latencies

הטוקנים לא מאוד רגישים למומחים שלהם מעבר למעט מומחים (מ-top-k) בעלי ציונים הגבוהים ביותר. כלומר ניתן ״להפעיל רק המומחים״ בלי פגיעה משמעתית בביצועים

לא כל הטוקנים הם שווי ערך כלומר יש טוקנים רגישים יותר לשימוש בחלק מהמומחים שלהם ויש כאלו שפחות. המחברים טוענים שניתן להסיק את רמת הרגישות של הטוקן מציוני רשת הניתוב עבורו

השלב של prefill (עיבוד פרומפט) רגיש יותר להחלפת המומחים משלב ה-decode (גנרוט)

הרגישות להחלפת המומחים משתנה בין שכבות המודל כאשר השכבות האמצעיות הן הרגישות ביותר לזה

המחברים מציעים לנצל את אובזרבציות אלו בצורה הבאה (יש כמה וריאציות, אתאר את עיקרי השיטה)

משתמשים בכל המומחים בשלב ה- prefill (שהוא compute-bound)

מזהים טוקנים רגישים ופחות רגישים (low and high confidence) בבאץ'. לאחר מכן מפלטרים את המומחים של הטוקנים הפחות רגישים

בוחרים את המומחים שהם הכי בשימוש עבור הבאץ' ומפלטרים את השאר

מפעילים רק את המומחים שנותרו עבור כל הטוקנים (top-k). אופציה שניה (פחות פוגעת בביצועים) - היא להפעיל את כל המומחים עבור טוקנים רגישים ורק את אלו שנותרו עבור טוקנים פחות רגישים

שיטה זו מאפשרת להגדיל צפיפות חישובות עבור שלב ה-decode ולעשות אותו פחות memory-bound בלי פגיעה משמעותית בביצועים.

https://arxiv.org/abs/2411.08982