⚡️🚀המאמר היומי של מייק 15.06.24:⚡️🚀

MEDUSA: Simple LLM Inference Acceleration Framework with Multiple Decoding Heads

ב 3 הסקירות האחרונות ראינו כמה שיטות איטרטיביות מקבילות, מבוססות על שיטות יאקובי ו- Gauss-Seidel, המנסות להאיץ את מהירות גנרוט הטקסט (decoding) של מודלי שפה. היום נסקור קצרות מאמר המציע גישה אחרת לאותה הבעיה, שגם מבצעת גנרוט מקבילי של טקסט אבל בשיטה 'טיפה' אחרת.

בגדול המאמר מציע להוסיף ולאמן כמה ״ראשים״ (שכבה לינארית עם סופטמקס) למודל שפה מאומן. מטרתה של כל ראש כזה היא לחזות טקסט לא החל מהטוקן הבא אלא להתחיל לחזות מהטוקן ה-k אחרי הפרומפט (או הטוקן האחרון שנחזה). כלומר בהינתן פרומפט באורך 10 טוקנים הראש מסדר 3 מגנרט טוקנים החל מהטוקן ה-14 בזמן שמודל שפה רגיל חוזה(מגנרט) החל מהטוקן ה-11. הראשים האלו מחוברים לשכבה האחרונה (לפני שכבת החיזוי) של מודל שפה. כלומר הם מפעילים טרנספורמציה לינארית על ייצוג(תלוי קונטקסט) הטוקן המופק על ידי מודל שפה.

המחברים מציעים שתי דרכים לאמן מודל שפה עם הראשים האלו. הדרך הראשונה היא לאמן רק את הראשים כאשר מודל השפה עצמו נותר מוקפא. הדרך השנייה היא לעשות פיין טיון של מודל שפה מאומן (עם LoRa כמובן). במקרה השני הם משלבים את הלוס הסטנדרטי של מודלי שפה עם זה של הראשים האחרים.

באינפרנס המחברים לוקחים את החיזויים מהראשים השונים (כמה טוקנים החל מטוקן k לכל ראש) של הראשים השונים ומשלבים אותם בצורה דומה ל- beam search (כאן זה קצת יותר מורכב ונקרא tree-search) כדי לקבל את כמה סדרות של טוקנים (המועמדות) שמהן נבנה החיזוי הסופי של מודל שפה. כדי לבחור את התת-סדרות של טוקנים ״הטובות ביותר״ ביותר הם עושים משהו דומה למה שנעשה ב-speculative decoding קלאסי (טיפה יותר מורכב משם ו-rejection sampling בעניין).

אז מה הרווח כאן אתם שואלים? שהראשים מופעלים באופן מקבילי ולפעמים בהפעלה אחת שלהם אנו חוזים כמה טוקנים ולא אחד כמו בגנרוט אוטורגרסיבי רגיל.

https://arxiv.org/pdf/2401.10774