⚡️🚀המאמר היומי של מייק 13.06.24:⚡️🚀

Break the Sequential Dependency of LLM Inference Using LOOKAHEAD DECODING

זוכרים את המאמר שסקרנו קצרות אתמול שהציע גישה איטרטיבית לפתרון מקבילי של מערכות משוואות לא לינאריות. אחת הדוגמאות של פתרון מערכות משוואות כאלו היא גנרוט טקסט ממודלי שפה כאשר כל טוקן נבחר בתור argmax של התפלגות הטוקן בהינתן הטוקנים הקודמים (המופק באמצעות השכבה האחרונה של מודל השפה).

יש בגדול שתי שיטות איטרטיביות שניתן לרתום אותן לדגימה יעילה יותר ממודלי שפה: יעקובי וגאוס-סיידל. שתי השיטות מתחילות מניחוש אקראי של כמה טוקנים בהינתן ההקשר ואז מאפטמים אותם על פתרון איטרטיבי של מערכת המשוואות עם argmax (ששקול לגנרוט). אפשר די בקלות לראות שבגלל שהמשוואות הן אוטורגרסיביות שיטות אלו לא יכולות להתכנס ביותר מ n איטרציות (מספר הטוקנים הנדגמים עם שיטה) ולפעמים אפשר להספיק פחות (נציין כי כל איטרציה דורשת קצת יותר משאבי החישוב).

הבעיה עם השימוש הנאיבי בשיטה הוא שהרווח הממוצע על פני דגימה אוטורגרסיבית סטנדרטית ממודלי שפה הוא לא גדול ועומד על פחות מ 1.1 האצת קצב גנרוט.

המאמר מציע שכלול לשיטה הנאיבית ומציע לשמור בזכרון את הטוקנים של כמה איטרציות האחרונות. במקרה אם והיא מוצאת בזכרון זה תת-סדרת טוקנים שבה הטוקן הראשון זהה לטוקן הראשון ״הנכון״ של האיטרציה(באיטרציה i טוקן i וקודמיו נחזים נכון) אנו לוקחים תת סדרה זו ומציבים אותו במקום מה שנחזה באיטרציה האחרונה.

זה מאפשר להקטין את כמות האיטרציות עוד טיפה

https://arxiv.org/pdf/2402.02057