⚡️🚀המאמר היומי של מייק 09.08.24: ⚡️🚀

Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters

בטח שמעתם על חוקי הסקיילינג של מודלי שפה. חוקים אלו מיועדים למציאת ״קונפיגורציה״ אופטימלית לאימון מודלי שפה. חוקי סקליינג מקשרים ערך של פונקציית לוס (ניתן להגדיר אותו בכמה אופנים) שניתן להשיגו עבור גודל מודל, גודל סט האימון וכמות משאבי החישוב (FLOps) המוקצית לאימון.

המאמר שואל האם ניתן לנסח חוקי סקיילנג דומים עבור האינפרנס, כלומר מה הביצועים המקסימליים שניתן להפיק בהינתן כמות משאבי חישוב נתונה. הרי יש כמה שיטות לבצע אינפרנס של מודל השפה ויש כמה פרמטרים חשובים של האינפרנס המשפיעים בצורה משמעותית על הביצועים. למשל יש שיטה הנקראת beam search שיוצרת בכל חיזוי של טוקן M סדרות טוקנים בעלי נראות (likelihood) הגבוהה ביותר. קיימות שיטות beam search עם מספר הסדרות השמורות לא קבוע ותלוי במספר הטוקן המגונרט.

יש שיטות איטרטיביות אחרות כמו במאמר "Consistency LLMs" שסקרתי לפני כמה שבועות. הוצעו גם שיטות שמשערכות את ״איכות״ התשובה המגונרטת (עם מודל מאומן נוסף) שמאפשר לבחור את התשובה הכי טובה מכמה תשובות מגונרטות (או להפסיק את יצירת התשובה אם רואים שהיא לא ״בכיוון). כל שיטה כזו דורשת משאבי חישוב שונים שתלויים גם בהייפרפרמטרים של השיטה.

מה השיטה העדיפה לרמת ביצועים אופטימלית בהינתן תקציב חישוב נתון (FLOps) - זו השאלה שהמאמר מנסה לענות עליה ויש תוצאות מעניינות (לדעתי)

https://arxiv.org/abs/2408.03314