המאמר היומי של מייק: 12.08.25  
Your LLM Knows the Future: Uncovering Its Multi-Token Prediction Potential

איך ניתן לגנרט טוקנים בצורה מקבילית אבל בלי מודלי שפה מבוססי דיפוזיה.

מאמר זה קורא תיגר לגנרוט אוטורגרסיבי של LLMs ומציע שיטה שמאמנת מודל לחזות כמה טוקנים בו זמנית כלומר MTP שזה Multiple Token Prediction. כאמור MTP מאומן לחזות כמה טוקנים בו זמנית להבדיל מ-NTP או Next Token Prediction שחוזה כל פעם טוקן יחיד. בנוסף הגישה המוצעת משלבת שימוש במה שנקרא פענוח ספקולטיבי או Speculative Decoding, נושא שהעברתי עליו הרצאות בכמה כנסים ומיטאפים לאחרונה. בנוסף יש גם שימוש בטכנית fine-tune של מודלים (בד״כ מבוססי טרנספורמרים) הנקראת LoRa שזה ראשי תיבות של Low Rank Adaptation.

אוקי, אז קודם כל המחברי מאמנים כמה ראשי decoding (למיטב הבנתי שכבה אחת בלבד) עבור כל טוקן שנחזה פרט לטוקן הבא שנחזה באופן סטנדרטי כמו ב-NTP. בשביל לחזות את הטוקן הבא המחברים משתמשים לא רק בייצוג הקונטקסטואלי שלו אלא גם בייצוג הלא קונטקסטואלי (ממילון האמבדינג) של הטוקן הקודם (שניהם משורשרים ומועברים דרך MLP בעל שתי שכבות).

בנוסף המאמר מאמן LoRA (מטריצות תוספות למשקולות של השכבות הלינאריות של הטרנספורמר) אבל משתמש בהם רק כדי לחזות את הטוקנים מעבר לטוקן הבא. במאמר שיטה זו נקראת Gated LoRA. שיטה זו ניתן לאמן בצורה מקבילית בדומה לאיך שאנו מאמנים NTP סטנדרטי.

הגישה האחרונה הנדונה במאמר היא פענוח ספקולטיבי או SD. בגדול SD הינה משפחה של טכניקות לשיפור מהירות הגנרוט השומרת על התפלגות הגנרוט כמו בגנרוט אוטורגרסיבי (כלומר עם NTP). בד״כ משמתמשים במודל חלש ומהיר יותר (לפעמים מודל כזה הוא חלק מהמודל שאנו רוצים לייעל) לגנרוט של כמה טוקנים ואז בודקים אותם עם מודל היעד באופן מקבילי. הטוקנים שיעברו את הבחינה בהצלחה מתקבלים וככה אנו יכול לקבל גנרוט מהיר יותר.

כאן במקום המודל הגדול משתמשים בגנרוט מקבילי של כמה טוקנים דרך MTP, מעבירים להם את הבחינה וככל שיותר טוקנים עוברים אותה, אנו מקבלים גנרוט מהיר יותר. בנוסף המאמר מציע להמשיך לגנרט עם MTP עוד k טוקנים (k הוא מספר הטוקנים המגונרטים עם MTP). עם כל k הטוקנים הראשוניים עוברים את הבדיקה אנו ממשיכים את תהליך הבדיקה עם k הטוקנים הבאים שעתיד לזרז את קצב הגנרוט עוד יותר.

מאמר קליל יחסית וכתוב היטב - מומלץ.

https://arxiv.org/abs/2507.11851