המאמר היומי של מייק: 14.04.25  
Draft Model Knows When to Stop: A Self-Verification Length Policy for Speculative Decoding

המאמר הזה משך את עיניי כבר בהסתכלות הראשונה בגלל צמד המילים ״Speculative Decoding״ או SD בקצרה שמאוד קרוב לליבי - אפילו הכנתי על זה מצגת די מקיפה שאני מציג אותה בפורומים שונים. SD מאפשר להגדיל את קצב גנרוט טקסט על ידי מודל שפה באמצעות שילוב מודל היעד עם מודל קטן מהיר יותר וכמובן יותר חלש ממודל היעד. המודל הקטן מייצר כמה טוקנים בצורה אוטורגרסיבית ומודל היעד חוזה מנצל טוקנים אלו כדי לחזות בו זמנית את הטוקנים הבאים שלו. זה מאפשר להגדיל את הקצב הדגימה של המודל הגדול בצורה ניכרת.

השיטה מנצלת את העובדה שצוואר הבקבוק של תהליך הגנרוט העברת דאטה בין הזכרונות של gpu (בפרט HBM הגדול ואיטי ו-SRAM הקן אך מהיר בחלק החישובי של ה-gpu). אז SD מבצע חיזוי מהיר עם המודל הקטן ואז החיזוי הבו זמני על ידי המודל הגדול עם הטוקנים שנחזו על ידי המודל הקטן. אבל יש שם קאץ' כמובן: כדי לקבל את אותה התפלגות הטוקנים עם המודל הגדול דרך ניצול הטוקנים של המודל הקטן יש צורך בלבצע סוג של rejection sampling או RS.

אזכיר ש-RS מאפשר לדגום מהתפלגות קלה לדגימה f כדי ליצור מדגם הדגום מהתפלגות אחרת g שקשה לדגום ממנה בצורה ישירה. אז אנו דוגמים נקודה x מ-f אז מקבלים את הדגימה בהסתברות השווה ליחס בין (f(x ל- (g(x (אם יחס זה גדול מ-1 הנקודה מתקבלת אוטומטית). ניתן להוכיח שנקודות שהנדגמות באופן זה מפולגת עם התפלגות הרצויה g.

אז במקרה שלנו (SD) אנו עושים משהו דומה עבור הטוקנים הנדגמים עם המודל קטן. במהלך השלב השני (דגימה בו זמנית מהמודל הגדול) עבור כל טוקן הנגדם מהמודל הקטן אנו מחשבים את היחס בין ההסתברויות של המודלים ואנו ״מקבלים״ את הטוקנים של המודל הקטן בהסתברות השווה ליחס סיבה. אחרי שהטוקן הראשון של המודל הקטן ״סורב״ (rejected) המודל הגדול מגנרט טוקן הבא עם המודל הגדול ואז המודל הקטן שוב מופעל לגנרט את הטוקנים הבאים. ד״א גם הטוקנים שמתקבלים מגונרטים עם עם התפלגות המחושבת משתי ההתפלגויות של הטוקן (של המודל הקטן ושל הגדול).

כמו שכבר הצלחתם להבין ״שליטה״ ב acceptance rate של טוקנים של המודל הקטן היא מאוד חשובה - באידיאל אנו רוצים לדגום מהמודל הקטן רק את הטוקנים שיתקבלו. המאמר מציע שיטה לשפר את ה-acceptance rate. המאמר מראה שהממוצע של acceptance rate (זה די קל) שווה להפרש בין 1 למה שנקרא total variation distance או TBD בקצרה בין ההתפלגויות של שני המודלים (המותנות בהקשר). ולמזלנו עומד לרשותנו אי שוויון לא ידוע במיוחד שמאפשר לחסום TBD מלמטה עם הפרש בין קרוס-אנטרופי בין התפלגויותם של שני המודלים (עבור טוקן נתון בהינתן הקשרו) לאנטרופיה של טוקן של המודל הקטן.

אבל כמובן שאנו לא יכולים לחשב את הקרוס אנטרופי בין התפלגויות אלו בשלב דגימה מהמודל הגדול עבור כל הטוקנים כי אנו דוגמים כל הטוקנים ממנו בו זמנית ולא יודעים מראש התפלגות מותנית של כל טוקן של המודל הגדול. אז המאמר ״משערך״ את הקרוס אנטרופי הזה על זמן מדגם די גדול דרך קבוע (קצת גדול מ 1) מוכפל באנטרופיה של הטוקן של המודל הקטן. אחרי שיש לנו את הקרוס-אנטרופי אנו יכולים לשערך את ה-acceptance rate עבור כל טוקן של המודל הקטן לפני הדגימה מהמודל הגדול. זה מאפשר לנו לקבוע את מספר הטוקנים מהמודל הקטן שעבורם תתבצע דגימה בו זמנית מהמודל הגדול - פשוט בוחרים טוקנים עד שה-acceptance rate המשוערך גבוה מאיזה סף.

רעיון נחמד אבל בחירת הקבוע בשלב האחרון לדעתי לא אופטימלית ואני מקווה שבקרוב ייצאו מחקרים המשפרים את ההיבט הזה של השיטה המוצעת.

https://arxiv.org/abs/2411.18462