⚡️🚀המאמר היומי של מייק 14.06.24:⚡️🚀

CLLMs: Consistency Large Language Models

בשתי הסקירות הקודמות(כדאי שתעברו עליהם כי נתתי שם קצת הסברים) דיברנו על שיטות איטרטיביות מקבילות לדגימה ממודלי שפה. השיטות האלו מבוססות על שיטות יאקובי או (Gauss-Seidel (GS. השיטות האלו מתחילות מכמות מסוימת n של טוקנים שנדגמים באקראי (או בצורה קצת יותר מושכלת) ואז מעדכנים טוקנים אלו בבת אחת באיטרציות עד שתנאי עצירה מתקיים(התכנסות). תנאי העצירה כאן הוא בד״כ שוויון בין הפלטים של איטרציות עוקבות.

מובן שאנו מעוניינים לסיים את התהליך במשמעות פחות איטרציות ממספר הטוקנים שאנו חוזים בו זמנית (ד״א ניתן להראות נדרשות לכל היותר ח איטרציות עד ההתכנסות).

שימו לב שמהלך האימון של מודלי שפה מותאם לשיטת הדגימה האוטו-רגרסיביות כאשר בוחרים טוקן בעל הסתסברות הגבוה ביותר ביהנתן הטוקנים הקודמים. אולם עכשיו אנו דוגמים בצורה אחרת ואולי ניתן להתחשב בזה במהלך האימון. כלומר במהלך האימון אשכרה דוגמים עם השיטה הזו (השילוב של יאקובי ו- GS).

וזה בדיוק מה שנסקור אותו היום עושה. המחברים מוסיפים עוד איבר ללוס הרגיל של מודלי שפה (הממקסם את הנראות המירבית של הדאטה). מטרת האיבר הזה היא לגרום למזעור של מספר האיטרציות עד להתכנסות של הדגימה האיטרטיבית.

המחברים בחנו שתי אופציות לאיבר הזה:

מזעור של מרחק (KL הפוך לדעתי אך לא צללתי לעומק) בין התפלגויות הטוקנים בנקודת ההתכנסות לבין התפלגויות טוקנים במהלך הדגימה האיטרטיבית (דוגמים האיטרציות באקראי).

מזעור מרחק בין התפלגויות הטוקנים באיטרציות עוקבות.

ואם חשבתם שיש דמיון בין השיטה הזו לבין המאמר של איליה סלוצקבר ושותפיו "Consistency Models" - אכן הוא קיים ואני אצלול בו בקרוב.

https://arxiv.org/abs/2403.00835