דגימה בקבוצות

שימוש יעיל במודלי שפה סיבתיים - Causal Language Models

מתחרים

יוני קרמר

ביה"ס

תיכון עירוני ד׳ ע"ש אהרון קציר, תל אביב

מורה מלווה

גב׳ לימור שיאון

מנחה

מר עידו גודיס

מודל שפה סיבתי

הוא אלגוריתם חיזוי הטקסטים המתקדם ביותר בעולם.

הוא מתאים לכל משימות הבינה המלאכותית בהן הקלט והפלט הם טקסטים. לדוגמה: עזרה אישית, מעקב אחר הוראות כתובות, תרגום, סיכום, מענה על שאלות, תכנות, כתיבה יצירתית.

מודלי השפה המוכרים ביותר הם chat GPT ו GPT 3.

האלגוריתם האוטו-רגרסיבי

תחילה, האלגוריתם קולט טקסט מהמשתמש.

האלגוריתם מעביר את הטקסט מהמשתמש ישירות למודל שפה סיבתי. מודל השפה הסיבתי מחזיר תחזית למילה הבאה בטקסט.

לאחר מכן, מעבירים למודל את הקלט ואת המילה שהוא חזה בשלב הקודם והמודל חוזה מה תהיה המילה השנייה בפלט. התהליך חוזר על עצמו n פעמים כך **שכדי ליצור טקסט באורך n מילים עלינו להשתמש במודל** שפה סיבתי **n פעמים.**

בעיית הטעות הנגררת

כל מודל בינה מלאכותית בהכרח עושה טעויות, לא משנה כמה הוא טוב.

באלגוריתם האוטו רגרסיבי, החל מהמילה השנייה בפלט, המודל מסתמך על מילים שהוא יצר על מנת ליצור עוד מילים. כאשר המודל טועה בתחזית לאחת המילים, כל המילים שלאחריה יהיו מבוססות על אותה טעות. דבר זה הופך את **הטקסטים שנוצרו על ידי האלגוריתם האוטו-רגרסיבי למלאי טעויות ולא אמינים,** במיוחד כאשר הטקסטים ארוכים.

בעיית עלויות המחשוב

האלגוריתם האוטו-רגרסיבי הוא בזבזני ואיטי מכיוון שהוא דורש שימושים רבים במודל שפה סיבתי ולכן הרצתו על מחשבי היא מאוד יקרה.

לדוגמה: עלות השימוש במודל chat GPT מוערכת **במעל שלושה מיליון דולר** בחודש.

שאלות המחקר

האם ניתן ליצור טקסט באורך n מילים בעזרת פחות מ n שימושים על ידי מודל שפה סיבתי?

האם ניתן למנוע את בעיית הטעות הנגררת?

תגלית המחקר

כאשר נותנים למודל רצף מילים המתחיל בקלט של המשתמש ואחריו n-1 מילים שהוא לא מכיר, המודל יחזה את n המילים הבאות בטקסט, ולא רק את המילה הבאה.

הפתרון - דגימה בקבוצות

אלגוריתם היוצר טקסט בכל אורך בעזרת שימוש אחד בלבד במודל שפה סיבתי.

באלגוריתם זה, המילים שבאמצע הפלט לא תלויות במילים שהאלגוריתם חזה בשלבים קודמים אלא רק בקלט של המשתמש. ככה **דגימה בקבוצות פותרת את בעיית הטעות הנגררת.**

כאשר m הוא מספר המילים בקלט ו n הוא מספר המילים בטקסט שהמודל יוצר, סיבוכיות זמן הריצה של דגימה בקבוצות היא:

$$O(n^2 + m^2)$$

זאת לעומת לעומת האלגוריתם האוטו-רגרסיבי שסיבוכיות זמן הריצה שלו היא:

$$O(n^3 + nm^2)$$

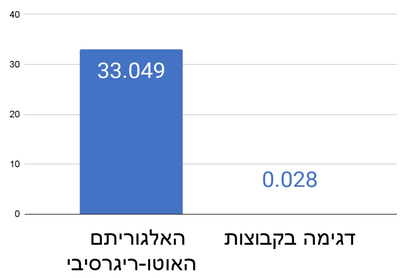
מה שאומר ש**דגימה בקבוצות יעילה יותר מהאלגוריתם האוטו-רגרסיבי** - במיוחד עבור טקסטים ארוכים.

ניסויים ותוצאות

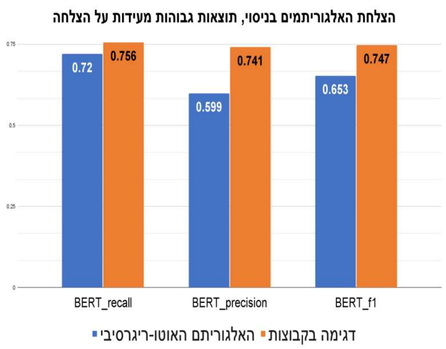
העבודה כללה ניסויים בהם שני האלגוריתמים תרגמו עשרות אלפי הרצאות TED בניסויים נמדדו זמן הריצה והצלחת התרגום של כל אלגוריתם.

זמן ריצה בשעות

זמן ריצה קטן מעיד על אלגוריתם יעיל



הרצת דגימה בקבוצות מהירה וזולה פי 1182.33 מהרצת האלגוריתם האוטו-רגרסיבי.



דגימה בקבוצות מצליחה בתרגום ב 5-24% **יותר מהאלגוריתם האוטו-רגרסיבי.**

מסקנות המחקר

- **דגימה בקבוצות עדיפה על האלגוריתם האוטו-רגרסיבי בכל אספקט**.
- היתרון של דגימה בקבוצות משמעותי יותר ככל שהאלגוריתם יוצר טקסטים ארוכים יותר.

חשיבות המחקר

האלגוריתם הוא פריצת דרך ה**משפרת משמעותית את איכות ואמינות הטקסטים** שנוצרים באמצעות מודלי שפה סיבתיים **ומורידה משמעותית את עלויות השימוש** בהם, גם לחברות וגם למשתמשי קצה.

שימוש באלגוריתם

ניתן להשתמש באלגוריתם דרך ספריית הקוד הפתוח grouped-sampling שפורסמה במסגרת העבודה.

נסו בעצמכם

המשך המחקר

- הוספת דגימה בקבוצות לספריות קוד פתוח.
- הרחבת התמיכה למודלי שפה סיבתיים נוספים.
- הרחבת הניסויים.

הנחיה מטעם התחרות

גב׳ ליטל שיריין

מר צביאל למברגר