⚡️🚀המאמר היומי של מייק -30.11.24: ⚡️🚀  
Fishing for Magikarp: Automatically detecting under-trained tokens in large language models

מאמר מעניין מבית חברת cohere, אחת החברות שמפתחות מודלי שפה foundational.

רקע:

המאמר חוקר סוגיה מעניינת של טוקנים לא מאומנים מספיק (under-trained) כלומר שלא נמצאים (או נמצאים בכמות מזערית) בדאטהסט אימון של המודל. סיבה אפשרית לקיום טוקנים כאלו נעוצה בעובדה שמילון הטוקנים לא תמיד נבנה על בסיס הדאטהסט שהמודל מאומן עליו.

מילון הטוקנים בנוי על דאטהסט קטן הרבה יותר מדאטהסט אימון העצום של המודל בשלב אימון מקדים (pretraining): הרי בניית מילון טוקנים עם אלגוריתמים קיימים על דאטהסט של עשרות טריליוני טוקנים איננה פיזיבילית חישובי. בגדול מאוד בוחרים תת-מילים ״השכיחים ביותר״ בדאטהסט (כולל סימני פיסוק וכדומה) לפי שיטה מסוימת (היום השיטה הפופולרית היא Byte-Pair Encoding או BPE, שיטה טוקניזציה נוספת נקראת WordPiece). וההבדלים בסט לטוקניזציה לבין זה לאימון המודל עלול להוביל ליצירת טוקנים מוזרים כמו \_TheNitrome.

הנוכחות של טוקנים שלא אומנו מספיק במודל מובילה למספר בעיות, כולל בזבוז קיבולת בטוקנייזר ופגיעה ביעילות המודל. בנוסף הם עלולים לגרום לפלט לא רצוי ולשבש אפליקציות downstream במיוחד בעידן שבו מודלי שפה משתמשים יותר ויותר בנתונים חיצוניים. כמובן שטוקנים כאלו ״מזמינים״ jailbreaks למיניהם. למרות שנעשתה עבודה מסוימת בזיהוי טוקנים בעייתיים אלה, עדיין חסרות שיטות אוטומטיות אמינות ומוסברות היטב שנבדקו על מגוון רחב של מודלים.

פרטי מחקר:

המאמר מציע לזיהוי טוקנים undertrained כאלו באמצעות טרנספורמציה מסוימת של מטריצה unembedding U כלומר המטריצה הממפה את ייצוג הטוקן לווקטור המכיל התפלגות הסתברותית עבור כל הטוקנים במילון.

המחברים מציינים כי פונקציית הלוס באימון ממוזערת כאשר ההסתברות של טוקנים שאינם בשימוש נחזית כ-0, ללא קשר לקלט, מה שגורם ללוגיטים שלהם להתכנס למינוס אינסוף. המאמר משער שהמודל יכול להשיג חיזוי כזו (לא תלוי בקלט) באמצעות חיסור של וקטור קבוע c משורות של U, מה שמוביל לתרומה שלילית קבועה לערכי הלוגיטים של טוקנים שאינם בשימוש.

המחברים מציעים את האלגוריתם הבא לזיהוי טוקנים undertrained:

מגדירים קבוצה S של טוקנים חשודים ל-undertrained (כלומר אינדקסים של שורות ב-U)

חשב את הרכיב העיקרי(principal component) הראשון c של U כאומדן לרכיב קבוע c. מכיוון שפונקציית הסופטמקס אינה משתנה להסטות קבועות, יש להקפיד להסיר רכיב קבוע כזה כדי למקסם את ההפרדה של טוקנים שאינם בשימוש.

הסר אותו כדי לקבל U′ = U − (c^T\*U)U.

חשב את וקטור האמבדינגס הממוצע של הטוקנים שאינם בשימוש u\_oov = U'\_i, i∈S.  
חשב את מרחקי הקוסיין (או מרחק L2) בין u\_oov לבין שאר השורות ב-U′.

הטוקנים שהמרחק הזה קטן יחסית לאחרים (באחוזון 2 נגיד) חשודים להיות טוקנים שאומנו מספיק. המאמר מצליב הסתברויות של הטוקנים החשודים ל-undertrained ומראה שהן קטנות מאוד ומשתנים לאט מאוד (בעיקר בגלל weight decay) באופן עקבי לאורך האימון (ללא קשר לקלט).

https://arxiv.org/abs/2405.05417