⚡️🚀המאמר היומי של מייק 27.07.24: ⚡️🚀

Data Mixture Inference: What do BPE Tokenizers Reveal about their Training Data?

אחרי שבוע שלא סקרתי עבודות על LLMs חוזר לנושא הזה היום עם סקירה של המאמר המציע התקפה מציע תקיפה על מודלי שפה מבוססת טוקנייזרים. ההתקפה מיועדת לגלות מה המשקל היחסי של דאטה מסוג מסוים (שפה, שפת תכנות וכדומה) בדאטהסט שעליו אומן מודל שפה. לא יודע עד כמה ההתקפה הזו חמורה אבל עושים זאת על סמך הטוקנייזר.

אם אתם זוכרים הטוקנייזרים נבנים על שילוב אותיות (מספרים, סימני פסוק הכדומה) הכי נפוצים בדאטהסט האימון. אם הדאטהסט מורכב מכמה שפות אז הטוקנים שייבחרו יכילו גם אותיות (ולפעמים מילים שלמות) מכמה שפות המופיעות בדאטהסט. בשיטת טוקניזציה מפורסמת הנקראת Byte Pair Encoding או BPE קודם כל מפצלים את הטקסט לבטים (bytes), מחפשים זוגות בתים הכי נפוץ בדאטהסט, מאחדים אותם לטוקן חדש וממשיכים את התהליך עד שמגיעים לגודל של מילון הטוקן (50k-100k היום במודלי שפה מודרניים).

אז המאמר מנצל את מבנה של אחגוריתם טוקניזציה כדי להציע אלגוריתם המבוסס על התכנות הלינארי למציאת אומדן למשקל יחסי של הדאטהסטים השונים בסט האימון של המודל.

https://arxiv.org/abs/2407.16607