המאמר היומי של מייק - 23.12.24:  
T-FREE: Tokenizer-Free Generative LLMs via Sparse Representations for Memory-Efficient Embeddings

שוב חוזרים לנושא הטוקנייזרים - מתברר שהוא יותר חם ממה שחשבתי. נתקלתי במאמר המעניין שיטה נוספת לטוקניזציה המבוססת על פונקציה האש n-grams. השיטה המוצעת באה להתמודד עם גודל העצום של המילון מלווה כל מודל שפה גדול (עשרות אלפי טוקנים לכל הפחות) וגם טוקנים דומים מאוד מבחינת האותיות האותיות שמצריכות אמבדינגים שונים שזה לא יעיל (לטענת המחברים).

המחברים מנסים שיטת טוקניזציה שה-encoding שלה המורכב משלבים הבאים:

פירוק של טקסט למה שהם קוראים טוקנים כאשר ב-T-FREE טוקנים אלו הם בעצם מילים

כל מילה מחולקת לסדרה של 3-grams לא זרים למשל מילה hello מיוצגת על ידי חמישה 3-grams הבאים: {\_He, Hel, ell, llo, lo\_}. מספר 3-grams בייצוג הזה בדרך כלל מספר n-grams במילה שווה למספר האותיות במילה

מקודדים כל 3-gram עם m פונקציות האש שכל אחת מהם מקבלת v ערכים אפשריים כאשר v הינו אחד הייפר-הפרמטרים של השיטה.

כך כל מילה מקודדת על ידי n\*m מספרים בין 0 ל-v כאשר n הינו אורך המילה (מספר אותיות). ייצוג המילה הוא ממוצע (ועיגול) של כל nm ערכים האלו.

כל ערך בין 0 ל-v מקודד על ידי וקטור נלמד כאשר v וקטורים אלו למעשה מהווים את המילון של השיטה

שלב האימון והפענוח (כלומר גנרוט של מילים) נראים קצת יותר מורכבים. קודם כל באימון המטרה היא לחזות את nm האשים של 3-grams של המילה הבאה. כלומר במקום בעיית multi-class בפענוח של הטוקניזציה הרגילה (חיזוי של טוקן ממילון הטוקנים) יש לנו כאן בעיית multi-label כאשר אנו חוזים n\*m האשים. שימו לב ש n תלוי באורך המילה כלומר יש לנו מספר ״לייבלים״ שונה לפי אורך המילה.

הפענוח לא ממש ברור לי האמת. כאשר אנו רוצים לחזות את המילה הבאה אנו קודם כל מחשבים את כל ההאשים עבור כל המילים האפשריות (זה די הרבה כי לכל מילה יש גם את כל ההטיות שלה לכל הפחות ובנוסף מילים בעלות אורכים שונים מקודדים עם מספר n\*m שונה של האשים). לאחר מכן בוחרים את המילה המיוצגת על יד האשים בעלי ״ההסתברות הגבוהה ביותר״. נזכור שהמודל חוזה הסתברות של כל ערך של האש מ 1 עד v (גודל המילון) ולא לגמרי ברור איך נבחרת קבוצת האשים בעלת הסתברות הגבוהה ביותר.

בקיצור מאמר נחמד אבל לא ברור לי העניין עם הפענוח…

https://arxiv.org/abs/2406.19223