המאמר היומי של מייק - 19.12.24:  
Large Concept Models: Language Modeling in a Sentence Representation Space

מאמר שני (גם הוצג ב-NeurIPS 2024) של מטה המציע קונספט די מהפכני למודלי שפה. במאמר שסקרתי אתמול הם הציע לוותר על הטוקנייזר הסטנדרטי במודלי שפה ובמאמר שנסקור היום הם הציע לוותר על חיזוי של טוקן הבא שהתרגלנו אליו כל כך ב-LLMs.

כמו שאתם בטח זוכרים LLMs מאומנים (באימון מקדים וב-SFT) באמצעות מקסום הנראות (likelihood) של דאטהסט אימון D, כלומר מקסום של הסתברות גנרוט של D עם המודל המאומן. כדי לעשות את זה אנו ממקסמים (ביחס לפרמטרי מודל השפה שלנו) הסתברות של כל הפיסת דאטה. מכיוון שכל פיסת דאטה מורכב מטוקנים ניתן לבטא אותה באמצעות חוק בייס כמכפלה של הסתברויות מותנות שכל טוקנים בהינתן הטוקנים הקודמים (כלומר הקונטקסט). וככה אני מגיעים לחיזוי של טוקן בהינתן הקונטקסט גם אימון וגם כמובן באינפרנס.

המאמר מציין כי ״חשיבה טוקן טוקן״ אלא בקונספטים כאשר אנו בונים את הדיבור שלנו (תוך כדי הדיבור). המאמר מציע להטיל את הגישה הזו למודל שפה כאשר קונספט מוגדר בתור משפט. כלומר המחברים מציעים לאמן מודל לחזות את המשפט הבא במקום חיזוי טוקן הבא שאנו רגילים אליו במודלי שפה סטנדרטיים.

אבל איך נחזה משפט, הרי זה משהו דיסקרטי ועבור אורך די צנוע של המשפט מספר הערכים האפשריים שהוא יכול להיות הינו מעריכי והופך להיות גבוה מדי כדי לבצע את החיזוי בו (כלומר סופטמקס בגודל עצום). אז המאמר מעביר אותנו למישור הרציף ומציע לאמן מודל, שקיבל שם Large Concept Model או LCM לחיזוי אמבדינג של המשפט בהינתן האמבדינגס של המשפטים הקודמים לא בחלון הקונטקסט. המאמר בוחן כמה פונקציות לוס שהפשוטה מהם היא L2 בין האמבדינג ה-ground-truth לבין החזוי (יש עוד כמה מעניינים בפרק 2.4.1 במאמר).

הדרך הנוספת שהמאמר הציע לבנות את האמבדינג של המשפט הבא הוא אימון מודל דיפוזיה מותנה (רעיון יפה מאוד לטעמי) לחיזוי האמבדינג שלו.

האמבדינג נבנה על ידי מודל embedder שהוא נשאר קבוע במהלך האימון. בנוסף ל-embedder (שהוא encoder) יש לנו גם דקודר שהופך את הקונספט (האמבדינג שלו) לטקסט.

מאמר די יפה, כתוב די ברור רק קצת ארוך מדי לדעתי…

https://arxiv.org/abs/2412.08821