המאמר היומי של מייק - 27.01.25  
FineZip : Pushing the Limits of Large Language Models for Practical Lossless Text Compression

בחרתי את המאמר הזה לסקירה כי יש לי חיבה גדולה לכל מה שקשור לדחיסה - דחיסה של מודלים, דחיסה של דאטה או כל דחיסה שהיא :). המאמר מציע שיטה נחמדה לדחוס דאטה. אתם בטח יודעים שהמודלים שלנו יודעים לדחוס דאטה בצורה לא רעה עם הייצוג הלטנטי (אמבדינג) שהם מפיקים מהדאטה. אם אני לא טועה אנו יכולים לדחוס תמונה ברזולוציה גבוהה פי 100 עם האמבדינג שלו. אבל הדחיסה הזו היא לא lossless. אכן ניתן לשחזר את התמונה מהאמבדינג שלה כך שהעין האנושית לא תבחין שום הבדל בין התמונה המשוחזרת לבין המקורית, אבל הן לא בהכרח ייצאו זהות. במקרה של טקסט זה יכול להיות קצת בעייתי כי אנו רוצים לשחזרו כמו שהוא.

המאמר המסוקר לעמות זאת מציע שיטה לדחיסה טקסט כך שניתן יהיה לשחזרו במדויק. השיטה המוצעת היא די פשוטה ואינטואיטיבית. הרי איך מודל שפה מגנרט טקסט - השכבה האחרונה שלו פולטת התפלגות מעל מרחב הטוקנים והטוקן נדגם מההתפלגות הזו (יש כמה שיטות). אנו יכולים להיעזר בהתפלגות זו כדי לדחוס את הטקסט שלנו. למשל כמו שהוצע במאמר LLMZip אני יכולים לקודד כל טוקן (בהנתן הקשר לפניו) על יד ראנק של ההסתברות שלו בהתפלגות עבור הטוקן הזה. ראנק זה בעצם המיקום של הטוקן ברשימת הטוקנים הממוינת (בסדר יורד) לפי הסתברות שלו בהתפלגות הטוקן הזה.

אם מודל השפה שאנו משתמשים בו הוא מאוד חזק ראנק זה יהיה קרוב ל- 1(או 2, 3 אבל לא 1000). וידוע כי סדרות כאלו ניתן לדחוס בצורה מאוד יעילה (קצב דחיסה גבוה). אז LLMZip הציע לטייב מודל שפה לטקסט שדוחסים אותו שזה לא פרקטי כי לכל טקסט צריך לשמור מודל משלו וגם האימון כבד. האממר המסוקר מציע להשתמש ב-LoRA (או PEFT אחר) לדחיסה כך שנצטרך לשמור גם את מטריצות התוספות (או adapters) לכל טקסט עם מודל שפה אחד לכולם.

עדיין זה לא מאוד פרקטי אבל מבחינה רעיונית די מעניין…

https://arxiv.org/abs/2409.17141