⚡️🚀המאמר היומי של מייק 20.08.24: ⚡️🚀

JPEG-LM: LLMs as Image Generators with Canonical Codec Representations

המאמר הזה תפס את עיניי כי מילה "jpeg" הופיע בשמו. למרות שלא יצא לי לעבוד בתחום של דחיסת דאטה אני מאוד אוהב את הנושא המרתק הזה. בנוסף המאמר הזה מדבר על מודל VQ-VAE שהיה די פופולרי לפני שמודלי דיפוזיה השתלטו לנו לחלוטין על GenAI בראייה הממוחשבת.

אוקיי, אז כל זה קשור? קודם כל jpeg זו גישה ידועה לדחיסת תמונות. המאמר גם מדבר על AVC/H.264 שהיא גישה לדחיסת וידאו המתבססת על עקרונות דומים לאלו של jpeg. בגדול jpeg עובד בצורה הבאה:

מחלקים תמונות לפאצ'ים באותו הגודל ועושים לכל אחד DCT - Discrete Cosine Transform (כמו התמרת פוריה ללא החלק המדומה).

מבצעים קווינטוט של מקדמים DCT לכל פאץ' כאשר המקדמים לתדרים גבוהים "נחתכים" בצורה רצינית יותר

משתמשים בקידוד run length וגם בקידוד האפמן כדי לדחוס את כל המקדמים המקונטטים של הפאצ'ים.

אוקיי, עכשיו נרענן לכם מזה VQ-VAE. קודם כל VAE זה מודל גנרטיבי שלומד לגנרט דאטה מהייצוג הלטנטי שלו (במימד נמוך). VAE מורכב מהאנקודר מהדקודר שהראשון בהם מאומן להפיק ייצוג של דאטה במימד נמוך והדקודר משחזר את הדאטה ממנו. VAE מאומן בצורה המשרה התפלגות נתונה (בד״כ גאוסית) על המרחב הלטנטי וזה מאפשר לגנרט דאטה חדש באמצעות הדקודר מווקטור הדגום מהתפלגות זו.

VQ-VAE היא שכלול של VAE כאשר הוא מאומן לגנרט תמונה בצורה סדרתית (מפאצ'ים/טוקנים ויזואליים) כאשר כל פאץ מיוצג על ידי וקטור (לטנטי) מהמילון שנלמד גם כן. כלומר התמונה נבנית פאץ'-פאץ' כאשר כל פאץ' (כלומר וקטור מהמיליון שמייצג אותו) נדגם בהינתן כל פאצ'ים שכבר גונרטו. זה בטח מזכיר לכם מודל שפה שמגנרט טוקנים בדיוק באותה צורה.

VQ-VAE מאומן בשני שלבים: בראשון מאמנים את האנקודר, המילון והדקורד (המשחזר פאצ'ים מהווקטורים במילון) ובשלב השני מאמנים מודל לחזות טוקן ויזואלי הבא בהינתן הטוקנים שכבר נוצרו.

המחברים שילבו את הרעיונות האלו (חלקית) ואימנו מודל שיודע לחזות ייצוג jpeg או avc בצורה סדרתית. אבל מה הטוקנים כאן? בדומה למודלי שפה המחברים השתמשו ב-BPE או byte-pair encoding (עם שפצורים קלים). מכאן המחברים בנו מודל היודע לרנרט ייצוג jpeg של התמונה שניתן להפוך אותו לתמונה די בקלות.

רעיון די חמוד אבל יש לי הרגשה שכבר ראיתי רעיונות דומים בעבר…

https://www.arxiv.org/abs/2408.08459