⚡️🚀המאמר היומי של מייק -02.12.24: ⚡️🚀  
Autoregressive Model Beats Diffusion: Llama for Scalable Image Generation

היסטוריה:

סקירה היום אעשה חזרה קטנה בזמן (מבחינתי) ואסקור מאמר בנושא הראיה הממוחשבת. פעם הייתי סוקר אותם בתדירות גבוהה יותר אולם לאחרונה רוב המאמרים שאני סוקר שייכים לדומיין הטקסטואלי כלומר NLP. לא אגלה לכם סוד אם אגיד לכם שהיום מודלי דיפוזיה (לרוב לטנטיים) די השתלטו על תחום גנרוט דאטה ויזואלי (כלומר תמונות ווידאו).

אולם לפני 3-4 שנים המצב בדומיין הויזואלי (בחלקו הגנרטיבי) היה די שונה. היו בו גם VAE שזה Variational AutoEncoders, גם זרימות מנורמלות (Normalized Flows) אבל מי ששלט בו באופן די מוחלט היה כמובן GANs כלומר Generative Adversarial Networks. וכמובן היו שילובים די מעניינים של השיטות הנ״ל שהגיעו לביצועים די יפים כמו VQGAN שזה שילוב של VAE ו-GAN.

רקע:

המאמר שנסקור היום מחזיר לחיים את VQGAN וטוען שניתן להגיע לתוצאות טובות יותר איתו (עם שכלול קל) ממודלי דיפוזיה גנרטיביים באותם הגדלים (= מספר פרמטרים). זו הצהרה די חזקה שמצריכה להבין מה המחברים שכללו ב-VQGAN שהוצע לפני 4 שנים.   
  
קודם כל אסביר בקצרה איך עובד VQGAN (סקרתי אותו בעבר הרחוק בהרחבה) אז תוכלי לקפוץ לשם להסברים מפורטים יותר. בגדול VQGAN מורכב מאנקודר שמטרתו לקודד (במרחב הלטנטי) את הפאצ'ים של תמונה, codebook, המורכב ומספר גדול של וקטורים המקודדים את הפאצ'ים האלו ודקודר שלמעשה הופך את ייצוגי פאצ'ים אלו (וקטורים) לפאצים המרכיבים תמונה.

אחרי הקידוד של פאץ' על ידי האנקודר הווקטור הכי קרוב (לפי מרחק L2 לדעתי) נבחר מה-codebook והוא מוזרם לדקודר (יחד עם עם הוקטורים הפאצ'ים האחרים). האנקודר וה-codebook מאומנים להחזיר וקטורים כמה שיותר קרובים אחד לשני (יש שם stop-gradient גם) והדקודר מאומן לשחזר את התמונה (נבדק לכל פאץ' בנפרד וגם יחד) בצורה המיטבית (נמדד על ידי דמיון perceptual נקרא LPIPS וגם יש לוס של גאן בפנים עם הדיסקרימנטור).

מה המאמר עשה:

אבל איך נשתמש בכל לגנרוט? לאחר סיום אימון של VQGAN, לוקחים את כל הייצוגים הלטנטיים של התמונות מהדאטסט ומאמנים דקודר של הטרנספורמר לחזות ייצוג של פאץ' בהינתן הפאצ'ים הקודמים. ופה נכנסים לנו LLM שאנו כה אוהבים כי המחברים מאמנים אחד הלמות (LLAMA) למשימה הזו. הרי יש לנו מילון (codebook) כמו בשפה טבעית רק שבמקום הטוקנים הרגילים יש לנו טוקנים ויזואליים.

וזה עובד לא רע (לפי הבדיקות שהם עשו)...

https://arxiv.org/abs/2406.06525