⚡️🚀המאמר היומי של מייק 02.09.24: ⚡️🚀

Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Mode

היום נסקור מאמר על מודל מולטימודלי בצורה די מעניינת. המודל שאימנו במאמר יודע לגנרט גם תמונות וגם דאטה טקסטואלי ומהווה שילוב של מודל דיפוזיה ומודל שפה.

הייחודיות של המודל הזה מתבטאת בכך שהיא מגנרטת גם את הדאטה הטקסטואלי וגם הדאטה הויזואלי בצורה שאנו מגנרטים טקסטים, כלומר טוקן אחרי טוקן (עבור תמונה זה למעשה טוקן ויזואלי או ייצוג של פאץ'). כלומר אם אנו צריכים לגנרט תמונה יחד עם תיאורה המלא המודל יגנרט את התיאור טוקן ואחרי טוקן (next token prediction או NTP) ואחרי שיסיים יגנרט את התמונה טוקן אחרי טוקן (בצורת NTP גם כן). זה די נחמד האמת.

המודל שהמאמר אימון מכיל 7 מיליארד פרמטרים שזה די צנוע למודלי שפה וגודל די סטנדרטי למודלי דיפוזיה גנרטיביים (המודל הגדול של stable diffusion מכיל בערך 8B פרמטרים). אבל כאן יש לנו מודל המשלב את שתי היכולות האלו (גנרוט תמונות וגנרוט טקסטים) באיכות די גבוהה.

אבל אין מאמנים את המודל הזה? בגדול בהינתן קלט שהוא ערבוב של תמונה וטקסט (למשל תמונה מעורבבת עם טקסט). עם הטקסט הכל פשוט, מזינים אותו טוקן אחרי טוקן. לפני כל תמונה מכניסים טוקן BOI המסמן את תחילת התמונה וכאשר כל הטוקנים הויזואליים של התמונה הוזנו מכניסים טוקן EOI לסימון סיום הזנת התמונה. כאמור טוקנים של תמונה זה טוקנים ויזואליים המהווים ייצוגים של פאצ'ים לאחר האנקודר (של VAE).

איך מאמנים את החיה הזו? לטקסט זה די ברור - מאמנים את המודל לחזות טוקן טוקן כמו ב-LLM עבור מילון טוקנים נתון. עבור התמונה מחלקים את התמונה לפאצים, מעבירים כל פאץ דרך האנקודר של VAE ומזינים את התוצאה כטוקן. הייצוגים של הטוקנים הויזואלים מועברים דרך שכבה לינארית או unet להורדת מימד. במהלך האימון לומדים להסיר רעש מהגרסאות המורעשות של ייצוגי הטוקנים הויזואליים.

בגנרוט המודל יוצר את התמונה פאץ' פאץ' מהרעש (אחרי הסרת הרעש וקטור הייצוג מוזן לדקודר של VAE כדי לחזות את הפאץ' עצמו). לאחרונה השיטה הזו ליצירת תמונה לא פופולרית במיוחד - רוב השיטות יוצרות את התמונה המלאה (מהייצוג הלטנטי שלה). וכמובן כל הטוקנים האלו מוזנים לטרנספורמר אחד גדול!

מאמר מעניין ומומלץ לקריאה!

https://arxiv.org/pdf/2408.11039