המאמר היומי של מייק - 22.01.25  
MONOFORMER: ONE TRANSFORMER FOR BOTH DIFFUSION AND AUTOREGRESSION

היום נעשה סקירה קצרה של מאמר די מעניין ששילב שני סוגים של מודלים, מודל שפה ומודל ויז'ן בטרנספורמר אחד. רוב המודלים מולטימודליים מורכבים מכמה מודלים שכל אחד מהם אחראי על הגנרוט של סוג דאטה אחד. למשל מודלי שפה ויזואליים בד״כ מורכבים משני מודלים: מודל שפה ומודל לגנרוט תמונות. המחברים מציעים ״לחבר״ את שני המודלים האלה למודל טרנספורמר אחד וזה נעשה בצורה די אינטואיטיבית.

קודם כל נציין כי שני המודלים האלו עובדים במרחב הטוקנים כאשר עבור מודלי שפה כל טוקן הוא חלק של מילה או מילה שלמה ואילו עבור מודל ויזואלי כל טוקן הוא פאץ' של תמונה. אז הניסיון לחבר אותם למודל אחד נראה די טבעי אך לא ברור האם ניתן לאמן אותו הטרנספורמר לגנרט שפה ותמונות כאחד.

המודל המוצע מגנרט שפה בדיוק כמו LLM רגיל, בצורה אוטורגרסיבית, כלומר, טוקן אחרי טוקן. אבל איך ניתן לשלב אותו עם מודל לגנרוט תמונות שכמובן מבוסס על מודלי דיפוזיה (בשנת 2025 זה האופציה הדיפולטית הרי). קודם כל צריך לזכור שמודל אוטורגרסיבי (לגנרוט שפה) עובד בצורה סיבתית (קוזלית), כלומר במהלך גנרוט טוקן n כל הטוקנים מאחוריו ממוסכים ולא משתתפים בגנרוט(משתמשים במסכה קוזלית). למודלי אנו צריכים מודל דו כיווני כי בזמן גנרוט פאץ' של תמונה כדאי מאוד להשתמש בכל הפאצ'ים האחרים.

בדיוק כך בנוי המודל המוצע - השפה מגונרטת עם מסכה קוזלית והתמונה מגונרטת עם כל הטוקנים (כולל הטוקנים של טקסט). דרך אגב הגישה הזו תעבוד גם לכיוון השני: כלומר בגנרוט של טקסט מתמונה (למשל למשימת captioning). אבל איך נדע לעבור ממצב ״קוזלי״ למצב ״דו-כיווני״. המחברים מציעים להשתמש בטוקן מסוים המסמן שממנו מתחיל גנרוט התמונה - הטוקן הזה אמור להיות מג'ונרט למשל למשימה יצירת תמונה מטקסט.

כמה מילים על הטרנספורמר לגנרוט תמונה. המאמר משתמש במודל דיפוזיה לטנטי כאשר המודל מאומן לבנות ייצוג לטנטי של תמונה מרעש (עבור כל פאץ). לאחר מכן כל הייצוגים (של הפאצ'ים) מועברים דרך הדקודר (מבוסס VAE) שבונה ממנו תמונה.

המודל מאומן עם הלוס שהוא סכום משוקלל של הלוסים הסטנדרטיים עבור המודלים המוזכרים: מודל שפה ומודל דיפוזיה. המאמר מצליח לגנרט תמונות די יפות….

https://arxiv.org/abs/2409.16280