המאמר היומי של מייק - 07.04.25  
JETFORMER: AN AUTOREGRESSIVE GENERATIVE MODEL OF RAW IMAGES AND TEXT

הסקירה של היום היא מאמר המשך(למרות שאין כאן VAE) של סקירתי האחרונה (מ 05.04.25). המאמר שנסקור היום מציע שיטה לאימון מודל מולטימודלי כאשר מודל אוטורגרסיבי אחד מאומן לשתי המודליות (תמונות וטקסט יחד). ברוב המודלים המולטי-מודליים יש אנקודרים שונים לטקסט ותמונה ולדעת מחברי המאמר זה עלול להוות בעיה (אני סוג של מבין את זה). אז המאמר מציע לאמן טרנספורמר אוטורגרסיבי לשתי המודליות יחד.

אז איך הדבר הזה עובד בעצם? המאמר מציע להשתמש במודל מאומן של זרימה מנורמלת (Normalized Flows or NF) לבניית ייצוג התמונה. מודל NF מאמן מיפוי הפיך ולכן lossless ממרחב הדאטה (תמונה) למרחב בעל התפלגות פשוטה (נגיד גאוסית סטנדרטית). בדרך כלל מיפוי זה נבנה על ידי הרכבה (composition) של כמה מיפוים פשוטים (נגיד על תת-קבוצה קטנה של מימדים) וכל המיפויים האלו מאמנים יחד כאשר המטרה היא למקסם את הנראות (likelihood) של הדאטה תחת המיפוי הזה. למעשה המחברים מאמנים NF עבור כל פאץ' בתמונה (ייצוג פאץ' נקרא טוקן ויזואלי).

אז המחברים מאמנים יחד מודל NF לייצוג תמונה יחד עם טרנספורמר אוטורגרסיבי לגנרוט תמונה וטקסט. כלומר בהינתן תיאור התמונה והתמונה עצמה (הסדר בהזנה של פיסות דאטה חשוב!) הטרנספורמר אוטורגרסיבי מאומן לפלוט את ייצוגי הטוקנים הויזואליים אחרי NF (שמאומנים יחד עם הטרנספורמר). כאשר תמונה מוזנת לפני התיאור שלה הטרנספורמר מאומן לשחזר את ייצוג הטוקנים הטקסטואליים. כמו בסקירה הקודמת (GIVT) המודל חוזה פרמטרים של ה-gaussian mixture עבור כל טוקן והייצוג נדגם משם.

המאמר גם מציע להעלות את הרובסטיות של ייצוגים המופקים על ידי המודל האוטורגרסיבי המאומן עם הרעשת דאטה(רק דאטה ויזואלי מורעש לפי הבנתי) מדורגת (סוג של למידת curriculum). בהתחלה מוסיפים לדאטה רעש חזק יותר כך שהמודל אוכל ללמוד את הפרטים ״הגסים״ של הדאטה ומורידים אותו במהלך האימון כך שהמודל ילמד גם את הפרטים העדינים יותר של הדאטה.

https://arxiv.org/abs/2411.19722